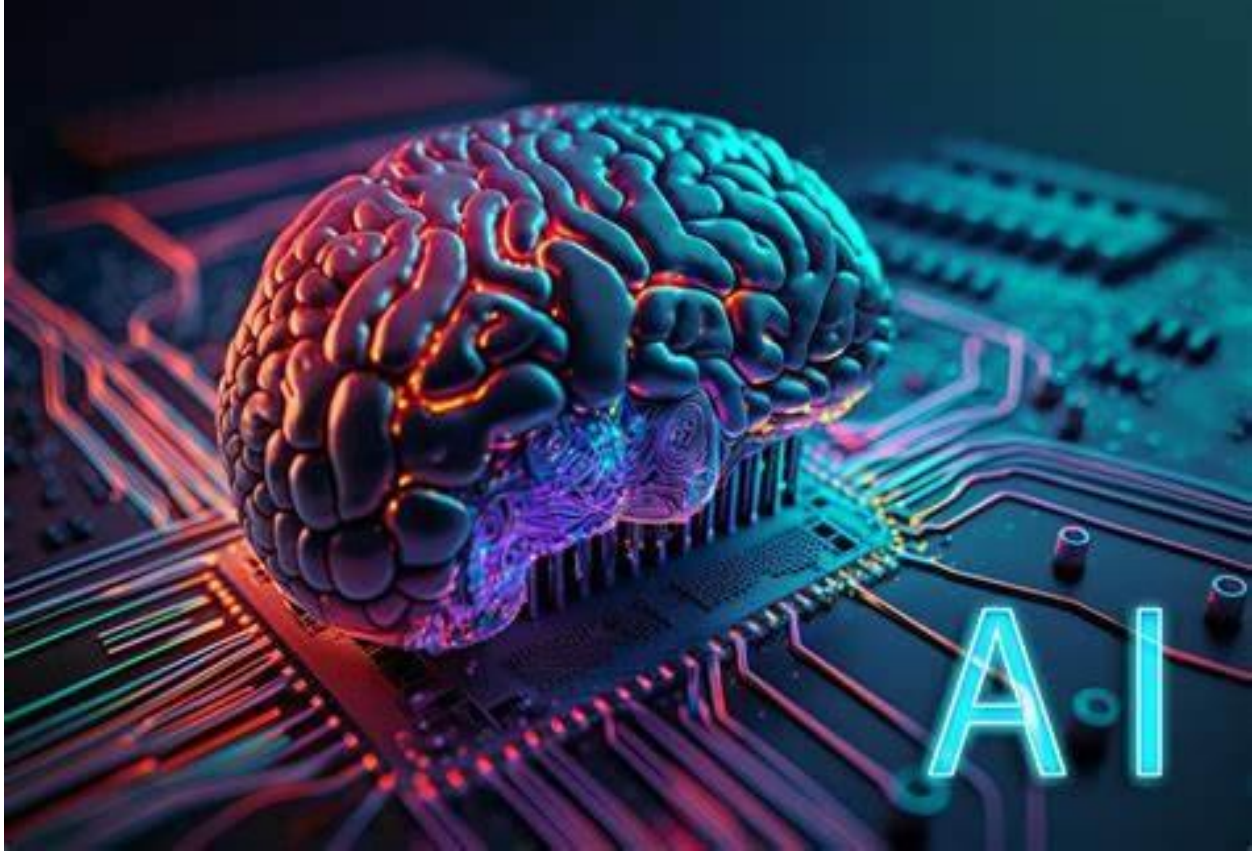


AI NEWS VOL 6



NATURAL LANGUAGE PROCESSING

Why Do Researchers Care About Small Language Models?

Larger models can pull off a wider variety of feats, but the reduced footprint of smaller models makes them attractive tools.

By pruning a language model, researchers can optimize it for a particular task.

Celsius Pictor for *Quanta Magazine*

Introduction

March 10, 2025

Large language models work well because they're so large. The latest models from OpenAI, Meta and DeepSeek use hundreds of billions of “parameters” — the adjustable knobs that determine connections among data and get tweaked during the training process. With more parameters, the models are better able to identify patterns and connections, which in turn makes them more powerful and accurate.

But this power comes at a cost. Training a model with hundreds of billions of parameters takes huge computational resources. To train its Gemini 1.0 Ultra model, for example, Google reportedly spent [\\$191 million](#)(). Large language models (LLMs) also require considerable computational power each time they answer a request, which makes them notorious energy hogs. A single query to ChatGPT [consumes about 10 times](#)() as much energy as a single Google search, according to the Electric Power Research Institute.

In response, some researchers are now thinking small. IBM, Google, Microsoft and OpenAI have all recently released small language models (SLMs) that use a few billion parameters — a fraction of their LLM counterparts.

Small models are not used as general-purpose tools like their larger cousins. But they can excel on specific, more narrowly defined tasks, such as summarizing conversations, answering patient questions as a health care chatbot and gathering data in smart devices. “For a lot of tasks, an 8 billion parameter model is actually pretty good,” said [Zico Kolter](#)(), a computer scientist at Carnegie Mellon University. They can also run on a laptop or cellphone, instead of a huge data center. (There’s no consensus on the exact definition of “small,” but the new models all max out around 10 billion parameters.)

To optimize the training process for these small models, researchers use a few tricks. Large models often scrape raw training data from the internet, and this data can be disorganized, messy and hard to process. But these large models can then generate a high-quality data set that can be used to train a small model. The approach, called knowledge distillation, gets the larger model to effectively pass on its training, like a teacher giving lessons to a student. “The reason [SLMs] get so good with such small models and such little data is that they use high-quality data instead of the messy stuff,” Kolter said.

Researchers have also explored ways to create small models by starting with large ones and trimming them down. One method, known as pruning, entails removing unnecessary or inefficient parts of a [neural network](#) — the sprawling web of connected data points that underlies a large model.

Pruning was inspired by a real-life neural network, the human brain, which gains efficiency by snipping connections between synapses as a person ages. Today’s pruning approaches trace back to [a 1989 paper](#)() in which the computer scientist Yann LeCun, now at Meta, argued that up to 90% of the parameters in a trained neural network could be removed without sacrificing efficiency. He called the method “optimal brain damage.” Pruning can help researchers fine-tune a small language model for a particular task or environment.

For researchers interested in how language models do the things they do, smaller models offer an inexpensive way to test novel ideas. And because they have fewer parameters than large models, their reasoning might be more transparent. “If you want to make a new model, you need to try things,” said [Leshem Choshen](#)(), a research scientist at the MIT-IBM Watson AI Lab. “Small models allow researchers to experiment with lower stakes.”

The big, expensive models, with their ever-increasing parameters, will remain useful for applications like generalized chatbots, image generators and [drug discovery](#)(). But for many users, a small, targeted model will work just as well, while being easier for researchers to train and build. “These efficient models can save money, time and compute,” Choshen said.

ARTIFICIAL INTELLIGENCE

The Unpredictable Abilities Emerging From Large AI Models

Large language models like ChatGPT are now big enough that they’ve started to display startling, unpredictable behaviors.

Paul Chaikin/Quanta Magazine

Introduction

March 16, 2023

That prompt was one of 204 tasks chosen last year to test the ability of various large language models (LLMs) — the computational engines behind AI chatbots such as ChatGPT. The simplest LLMs produced surreal responses. “The movie is a movie about a man who is a man who is a man,” one began. Medium-complexity models came closer, guessing *The Emoji Movie*. But the most complex model nailed it in one guess: *Finding Nemo*.

“Despite trying to expect surprises, I’m surprised at the things these models can do,” said [Ethan Dyer\(\)](#), a computer scientist at Google Research who helped organize the test. It’s surprising because these models supposedly have one directive: to accept a string of text as input and predict what comes next, over and over, based purely on statistics. Computer scientists anticipated that scaling up would boost performance on known tasks, but they didn’t expect the models to suddenly handle so many new, unpredictable ones.

Recent investigations like the one Dyer worked on have revealed that LLMs can produce hundreds of “emergent” abilities — tasks that big models can complete that smaller models can’t, many of which seem to have little to do with analyzing text. They range from multiplication to generating executable computer code to, apparently, decoding movies based on emojis. New analyses suggest that for some tasks and some models, there’s a threshold of complexity beyond which the functionality of the model skyrockets. (They also suggest a dark flip side: As they increase in complexity, some models reveal new biases and inaccuracies in their responses.)

Despite trying to expect surprises, I’m surprised at the things these models can do.

[Ethan Dyer, Google Research](#)

“That language models can do these sort of things was never discussed in any literature that I’m aware of,” said [Rishi Bommasani\(\)](#), a computer scientist at Stanford University. Last year, he helped compile [a list\(\)](#) of dozens of emergent behaviors, including several identified in Dyer’s project. That list continues to [grow\(\)](#).

Now, researchers are racing not only to identify additional emergent abilities but also to figure out why and how they occur at all — in essence, to try to predict unpredictability. Understanding emergence could reveal answers to deep questions around AI and machine learning in general, like whether

complex models are truly doing something new or just getting really good at statistics. It could also help researchers harness potential benefits and curtail emergent risks.

“We don’t know how to tell in which sort of application is the capability of harm going to arise, either smoothly or unpredictably,” said [Deep Ganguli\(\)](#), a computer scientist at the AI startup Anthropic.

The Emergence of Emergence

Biologists, physicists, ecologists and other scientists use the term “emergent” to describe self-organizing, collective behaviors that appear when a large collection of things acts as one. Combinations of lifeless atoms give rise to living cells; water molecules create waves; murmurations of starlings swoop through the sky in changing but identifiable patterns; cells make muscles move and hearts beat. Critically, emergent abilities show up in systems that involve lots of individual parts. But researchers have only recently been able to document these abilities in LLMs as those models have grown to enormous sizes.

Language models have been around for decades. Until about five years ago, the most powerful were based on what’s called a recurrent neural network. These essentially take a string of text and predict what the next word will be. What makes a model “recurrent” is that it learns from its own output: Its predictions feed back into the network to improve future performance.

In 2017, researchers at Google Brain introduced a new kind of architecture called a [transformer](#). While a recurrent network analyzes a sentence word by word, the transformer processes all the words at the same time. This means transformers can process big bodies of text in parallel.

It may very well be that the model has learned something fundamentally new and different that it didn’t have at a smaller size.

[Ellie Pavlick, Brown University](#)

Transformers enabled a rapid scaling up of the complexity of language models by increasing the number of parameters in the model, as well as other factors. The parameters can be thought of as connections between words, and models

improve by adjusting these connections as they churn through text during training. The more parameters in a model, the more accurately it can make connections, and the closer it comes to passably mimicking human language. As expected, a 2020 [analysis\(\)](#) by OpenAI researchers found that models improve in accuracy and ability as they scale up.

But the debut of LLMs also brought something truly unexpected. Lots of somethings. With the advent of models like GPT-3, which has 175 billion parameters — or Google’s PaLM, which can be scaled up to 540 billion — users began describing more and more emergent behaviors. One DeepMind engineer even [reported\(\)](#) being able to convince ChatGPT that it was a Linux terminal and getting it to run some simple mathematical code to compute the first 10 prime numbers. Remarkably, it could finish the task faster than the same code running on a real Linux machine.

As with the movie emoji task, researchers had no reason to think that a language model built to predict text would convincingly imitate a computer terminal. Many of these emergent behaviors illustrate “zero-shot” or “few-shot” learning, which describes an LLM’s ability to solve problems it has never — or rarely — seen before. This has been a long-time goal in artificial intelligence research, Ganguli said. Showing that GPT-3 could solve problems without any explicit training data in a zero-shot setting, he said, “led me to drop what I was doing and get more involved.”

He wasn’t alone. A raft of researchers, detecting the first hints that LLMs could reach beyond the constraints of their training data, are striving for a better grasp of what emergence looks like and how it happens. The first step was to thoroughly document it.

Ethan Dyer helped probe what kinds of unexpected abilities large language models were capable of, and what could bring them about.

Gabrielle Lurie

Beyond Imitation

In 2020, Dyer and others at Google Research predicted that LLMs would have transformative effects — but what those effects would be remained an open question. So they asked the research community to provide examples of difficult and diverse tasks to chart the outer limits of what an LLM could do. This effort was called the Beyond the Imitation Game Benchmark (BIG-bench) project, riffing on the name of Alan Turing’s “imitation game,” a test for

whether a computer could respond to questions in a convincingly human way. (This would later become known as the Turing test.) The group was especially interested in examples where LLMs suddenly attained new abilities that had been completely absent before.

“How we understand these sharp transitions is a great research question,” Dyer said.

As one would expect, on some tasks a model’s performance improved smoothly and predictably as complexity increased. And on other tasks, scaling up the number of parameters did not yield any improvement. But for about 5% of the tasks, the researchers found what they called “breakthroughs” — rapid, dramatic jumps in performance at some threshold scale. That threshold varied based on the task and model.

For example, models with relatively few parameters — only a few million — could not successfully complete three-digit addition or two-digit multiplication problems, but for tens of billions of parameters, accuracy spiked in some models. Similar jumps occurred for other tasks including decoding the International Phonetic Alphabet, unscrambling a word’s letters, identifying offensive content in paragraphs of Hinglish (a combination of Hindi and English), and generating a similar English equivalent of Kiswahili proverbs.

But the researchers quickly realized that a model’s complexity wasn’t the only driving factor. Some unexpected abilities could be coaxed out of smaller models with fewer parameters — or trained on smaller data sets — if the data was of sufficiently high quality. In addition, how a query was worded influenced the accuracy of the model’s response. When Dyer and his colleagues posed the movie emoji task using a multiple-choice format, for example, the accuracy improvement was less of a sudden jump and more of a gradual increase with more complexity. And last year, in a [paper presented at NeurIPS\(\)](#), the field’s flagship meeting, researchers at Google Brain showed how a model prompted to explain itself (a capacity called chain-of-thought reasoning) could correctly solve a math word problem, while the same model without that prompt could not.

You won’t know until you study the influence of scale what capabilities or limitations might arise.

Deep Ganguli, Anthropic

[Yi Tay\(\)](#), a scientist at Google Brain who worked on the systematic investigation of breakthroughs, points to recent work suggesting that chain-of-thought prompting changes the scaling curves and therefore the point where emergence occurs. In their NeurIPS paper, the Google researchers showed that using chain-of-thought prompts could elicit emergent behaviors not identified in the BIG-bench study. Such prompts, which ask the model to explain its reasoning, may help researchers begin to investigate why emergence occurs at all.

Recent findings like these suggest at least two possibilities for why emergence occurs, said [Ellie Pavlick\(\)](#), a computer scientist at Brown University who studies computational models of language. One is that, as suggested by comparisons to biological systems, larger models truly do gain new abilities spontaneously. “It may very well be that the model has learned something fundamentally new and different that it didn’t have at a smaller size,” she said. “That’s what we’re all hoping is the case, that there’s some fundamental shift that happens when models are scaled up.”

The other, less sensational possibility, she said, is that what appears to be emergent may instead be the culmination of an internal, statistics-driven process that works through chain-of-thought-type reasoning. Large LLMs may simply be learning heuristics that are out of reach for those with fewer parameters or lower-quality data.

But, she said, finding out which of those explanations is more likely hinges on a better understanding of how LLMs work at all. “Since we don’t know how they work under the hood, we can’t say which of those things is happening.”

Unpredictable Powers and Pitfalls

There is an obvious problem with asking these models to explain themselves: They are notorious liars. “We’re increasingly relying on these models to do basic work,” Ganguli said, “but I do not just trust these. I check their work.” As one of many amusing examples, in February Google introduced its AI chatbot, Bard. The blog post announcing the new tool shows Bard.

Emergence leads to unpredictability, and unpredictability — which seems to increase with scaling — makes it difficult for researchers to anticipate the consequences of widespread use.

To Teach Computers Math, Researchers Merge AI Approaches

FEBRUARY 15, 2023

“It’s hard to know in advance how these models will be used or deployed,” Ganguli said. “And to study emergent phenomena, you have to have a case in mind, and you won’t know until you study the influence of scale what capabilities or limitations might arise.”

In an [analysis of LLMs\(\)](#) released last June, researchers at Anthropic looked at whether the models would show certain types of racial or social biases, not unlike [those previously reported\(\)](#) in non-LLM-based algorithms used to predict which former criminals are likely to commit another crime. That study was inspired by an apparent paradox tied directly to emergence: As models improve their performance when scaling up, they may also increase the likelihood of unpredictable phenomena, including those that could potentially lead to bias or harm.

“Certain harmful behaviors kind of come up abruptly in some models,” Ganguli said. He points to [a recent analysis\(\)](#) of LLMs, known as the BBQ benchmark, which showed that social bias emerges with enormous numbers of parameters. “Larger models abruptly become more biased.” Failure to address that risk, he said, could jeopardize the subjects of these models.

But he offers a counterpoint: When the researchers simply told the model not to rely on stereotypes or social biases — literally by typing in those instructions — the model was less biased in its predictions and responses. This suggests that some emergent properties might also be used to reduce bias. [In a paper\(\)](#) released in February, the Anthropic team reported on a new “moral self-correction” mode, in which the user prompts the program to be helpful, honest and harmless.

Emergence, Ganguli said, reveals both surprising potential and unpredictable risk. Applications of these large LLMs are already proliferating, so a better understanding of that interplay will help harness the diversity of abilities of language models.

“We’re studying how people are actually using these systems,” Ganguli said. But those users are also tinkering, constantly. “We spend a lot of time just chatting with our models,” he said, “and that is actually where you start to get a good intuition about trust — or the lack thereof.”

QUANTIZED COLUMNS

What Does It Mean for AI to Understand?

It’s simple enough for AI to seem to comprehend data, but devising a true test of a machine’s knowledge has proved difficult.



Maggie Chiang for Quanta Magazine

Introduction

Remember IBM's Watson, the AI *Jeopardy!* champion? "Watson understands natural language with all its ambiguity and complexity." However, as we saw when Watson subsequently failed spectacularly in its quest to "[revolutionize medicine with artificial intelligence\(\)](#)," a veneer of linguistic facility is not the same as actually comprehending human language.

[Natural language understanding](#) has long been a major goal of AI research. At first, researchers tried to manually program everything a machine would need to make sense of news stories, fiction or anything else humans might write. This approach, as Watson showed, was futile — it's impossible to write down all the unwritten facts, rules and assumptions required for understanding text. More recently, a new paradigm has been established: Instead of building in explicit knowledge, we let machines learn to understand language on their own, simply by ingesting vast amounts of written text and learning to predict words. The result is what researchers call a language model. When based on large neural networks, like OpenAI's GPT-3, such models can generate uncannily humanlike [prose\(\)](#) (and [poetry\(\)](#)!) and seemingly perform sophisticated [linguistic reasoning\(\)](#).

But has GPT-3 — trained on text from thousands of websites, books and encyclopedias — transcended Watson's veneer? Does it really understand the language it generates and ostensibly reasons about? This is a topic of stark disagreement in the AI research community. Such discussions used to be the purview of philosophers, but in the past decade AI has burst out of its academic bubble into the real world, and its lack of understanding of that world can have real and sometimes devastating consequences. [In one study\(\)](#), IBM's Watson was found to propose "multiple examples of unsafe and incorrect treatment recommendations." Another study showed that Google's machine translation system [made significant errors\(\)](#) when used to translate medical instructions for non-English-speaking patients.

Quantized Columns

A regular column in which top researchers explore the process of discovery. This month's columnist, Melanie Mitchell, is the Davis Professor of Complexity at the Santa Fe Institute and the author of [Artificial Intelligence: A Guide for Thinking Humans](#).

How can we determine in practice whether a machine can understand? In 1950, the computing pioneer Alan Turing tried to answer this question with his famous “imitation game,” now called the Turing test. A machine and a human, both hidden from view, would compete to convince a human judge of their humanness using only conversation. If the judge couldn’t tell which one was the human, then, Turing asserted, we should consider the machine to be thinking — and, in effect, understanding.

Unfortunately, Turing underestimated the propensity of humans to be fooled by machines. Even simple chatbots, such as Joseph Weizenbaum’s 1960s ersatz psychotherapist [Eliza\(\)](#), have fooled people into believing they were conversing with an understanding being, even when they knew that their conversation partner was a machine.

In a [2012 paper\(\)](#), the computer scientists Hector Levesque, Ernest Davis and Leora Morgenstern proposed a more objective test, which they called the Winograd schema challenge. This test has since been adopted in the AI language community as one way, and perhaps the best way, to assess machine understanding — though as we’ll see, it is not perfect. A Winograd schema, named for the language researcher Terry Winograd, consists of a pair of sentences, differing by exactly one word, each followed by a question. Here are two examples:

Sentence 1: I poured water from the bottle into the cup until it was full.

Question: What was full, the bottle or the cup?

Sentence 2: I poured water from the bottle into the cup until it was empty.

Question: What was empty, the bottle or the cup?

Sentence 1: Joe’s uncle can still beat him at tennis, even though he is 30 years older.

Question: Who is older, Joe or Joe’s uncle?

Sentence 2: Joe’s uncle can still beat him at tennis, even though he is 30 years younger.

Question: Who is younger, Joe or Joe’s uncle?

Neural network language models have achieved about 97% accuracy on a particular set of Winograd schemas. This roughly equals human performance.

In each sentence pair, the one-word difference can change which thing or person a pronoun refers to. Answering these questions correctly seems to require commonsense understanding. Winograd schemas are designed precisely to test this kind of understanding, alleviating the Turing test’s

vulnerability to unreliable human judges or chatbot tricks. In particular, the authors designed a few hundred schemas that they believed were “Google-proof”: A machine shouldn’t be able to use a Google search (or anything like it) to answer the questions correctly.

These schemas were the subject of a competition held in 2016 in which the winning program was correct on only 58% of the sentences — hardly a better result than if it had guessed. Oren Etzioni, a leading AI researcher, [quipped\(\)](#), “When AI can’t determine what ‘it’ refers to in a sentence, it’s hard to believe that it will take over the world.”

However, the ability of AI programs to solve Winograd schemas rose quickly due to the advent of large neural network language models. A [2020 paper from OpenAI\(\)](#) reported that GPT-3 was correct on nearly 90% of the sentences in a benchmark set of Winograd schemas. Other language models have performed even better after training specifically on these tasks. At the time of this writing, neural network language models have achieved about 97% accuracy on a particular set of Winograd schemas that are part of an AI language-understanding competition known as [SuperGLUE\(\)](#). This accuracy roughly equals human performance. Does this mean that neural network language models have attained humanlike understanding?

Not necessarily. Despite the creators’ best efforts, those Winograd schemas were not actually Google-proof. These challenges, like many other current tests of AI language understanding, sometimes permit [shortcuts\(\)](#) that allow neural networks to perform well without understanding. For example, consider the sentences “The sports car passed the mail truck because it was going faster” and “The sports car passed the mail truck because it was going slower.” A language model trained on a huge corpus of English sentences will have absorbed the correlation between “sports car” and “fast,” and between “mail truck” and “slow,” and so it can answer correctly based on those correlations alone rather than by drawing on any understanding. It turns out that many of the Winograd schemas in the SuperGLUE competition allow for these kinds of statistical correlations.

[The Computer Scientist Training AI to Think With Analogies](#)

[Q & A](#)

The Computer Scientist Training AI to Think With Analogies

JULY 14, 2021

Rather than give up on the Winograd schemas as a test of understanding, a group of researchers from the Allen Institute for Artificial Intelligence decided instead to try to fix some of their problems. In 2019 they created [WinoGrande](#), a much larger set of Winograd schemas. Instead of several hundred examples, WinoGrande contains a whopping 44,000 sentences. To obtain that many examples, the researchers turned to Amazon Mechanical Turk, a popular platform for crowdsourcing work. Each (human) worker was asked to write several sentence pairs, with some constraints to ensure that the collection would contain diverse topics, though now the sentences in each pair could differ by more than one word.

The researchers then attempted to eliminate sentences that could allow statistical shortcuts by applying a relatively unsophisticated AI method to each sentence and discarding any that were too easily solved. As expected, the remaining sentences presented a much harder challenge for machines than the original Winograd schema collection. While humans still scored very high, neural network language models that had matched human performance on the original set scored much lower on the WinoGrande set. This new challenge seemed to redeem Winograd schemas as a test for commonsense understanding — as long as the sentences were carefully screened to ensure that they were Google-proof.

However, another surprise was in store. In the almost two years since the WinoGrande collection was published, neural network language models have grown ever larger, and the larger they get, the better they seem to score on this new challenge. At the time of this writing, the current best programs — which have been trained on terabytes of text and then further trained on thousands of WinoGrande examples — get close to [90% correct](#)() (humans get about 94% correct). This increase in performance is due almost entirely to the increased size of the neural network language models and their training data.

Understanding language requires understanding the world, and a machine exposed only to language cannot gain such an understanding.

Have these ever larger networks finally attained humanlike commonsense understanding? Again, it's not likely. The WinoGrande results come with some important caveats. For example, because the sentences relied on Amazon Mechanical Turk workers, the quality and coherence of the writing is quite

uneven. Also, the “unsophisticated” AI method used to weed out “non-Google-proof” sentences may have been too unsophisticated to spot all possible statistical shortcuts available to a huge neural network, and it only applied to individual sentences, so some of the remaining sentences ended up losing their “twin.” [One follow-up study](#)) showed that neural network language models tested on twin sentences only — and required to be correct on both — are much less accurate than humans, showing that the earlier 90% result is less significant than it seemed.

So, what to make of the Winograd saga? The main lesson is that it is often hard to determine from their performance on a given challenge if AI systems truly understand the language (or other data) that they process. We now know that neural networks often use statistical shortcuts — instead of actually demonstrating humanlike understanding — to obtain high performance on the Winograd schemas as well as many of the most popular [“general language understanding” benchmarks](#)).

The crux of the problem, in my view, is that understanding language requires understanding the world, and a machine exposed only to language cannot gain such an understanding. Consider what it means to understand “The sports car passed the mail truck because it was going slower.” You need to know what sports cars and mail trucks are, that cars can “pass” one another, and, at an even more basic level, that vehicles are objects that exist and interact in the world, driven by humans with their own agendas.

All this is knowledge that we humans take for granted, but it’s not built into machines or likely to be explicitly written down in any of a language model’s training text. Some cognitive scientists have argued that humans rely on innate, pre-linguistic [core knowledge](#)([opens a new tab](#)) of space, time and many other essential properties of the world in order to learn and understand language. If we want machines to similarly master human language, we will need to first endow them with the primordial principles humans are born with. And to assess machines’ understanding, we should start by assessing their grasp of these principles, which one might call [infant metaphysics](#).”

Training and evaluating machines for baby-level intelligence may seem like a giant step backward compared to the prodigious feats of AI systems like Watson and GPT-3. But if true and trustworthy understanding is the goal, this may be the only path to machines that can genuinely comprehend what “it” refers to in a sentence, and everything else that understanding “it” entails.

How Close Are Computers to Automating Mathematical Reasoning?

AI tools are shaping next-generation theorem provers and with them the relationship between math and man.

Introduction

August 27, 2020

In the 1970s, the late mathematician Paul Cohen, the only person to ever win a Fields Medal for work in mathematical logic, [reportedly](#) made a sweeping prediction that continues to excite and irritate mathematicians — that “at some unspecified future time, mathematicians would be replaced by computers.” Cohen, legendary for his daring methods in set theory, predicted that all of mathematics could be automated, including the writing of proofs. A proof is a step-by-step logical argument that verifies the truth of a [conjecture](#), or a mathematical proposition. (Once it’s proved, a conjecture becomes a theorem.) It both establishes the validity of a statement and explains why it’s true. A proof is strange, though. It’s abstract and untethered to material experience. “They’re this crazy contact between an imaginary, nonphysical world and biologically evolved creatures,” said the cognitive scientist [Simon DeDeo](#)() of Carnegie Mellon University, who studies mathematical certainty by analyzing the structure of proofs. “We did not evolve to do this.”

Computers are useful for big calculations, but proofs require something different. Conjectures arise from inductive reasoning — a kind of intuition about an interesting problem — and proofs generally follow deductive, step-by-step logic. They often require complicated creative thinking as well as the more laborious work of filling in the gaps, and machines can’t achieve this combination.

Computerized theorem provers can be broken down into two categories. Automated theorem provers, or ATPs, typically use brute-force methods to

crunch through big calculations. Interactive theorem provers, or ITPs, act as proof assistants that can verify the accuracy of an argument and check existing proofs for errors. But these two strategies, even when combined (as is the case with newer theorem provers), don't add up to automated reasoning.

Simon DeDeo of Carnegie Mellon helped show that people and machines seem to construct mathematical proofs in similar ways.

Courtesy of Simon DeDeo

Plus, the tools haven't been met with open arms, and the majority of mathematicians don't use or welcome them. "They're very controversial for mathematicians," DeDeo said. "Most of them don't like the idea."

A formidable open challenge in the field asks how much proof-making can actually be automated: Can a system generate an interesting conjecture and prove it in a way that people understand? A slew of recent advances from labs around the world suggests ways that artificial intelligence tools may answer that question. Josef Urban at the Czech Institute of Informatics, Robotics and Cybernetics in Prague is exploring a variety of approaches that use machine learning to boost the efficiency and performance of existing provers. In July, his group [reported](#)() a set of original conjectures and proofs generated and verified by machines. And in June, a group at Google Research led by [Christian Szegedy](#)) posted recent results from efforts to harness the strengths of natural language processing to make computer proofs more human-seeming in structure and explanation.

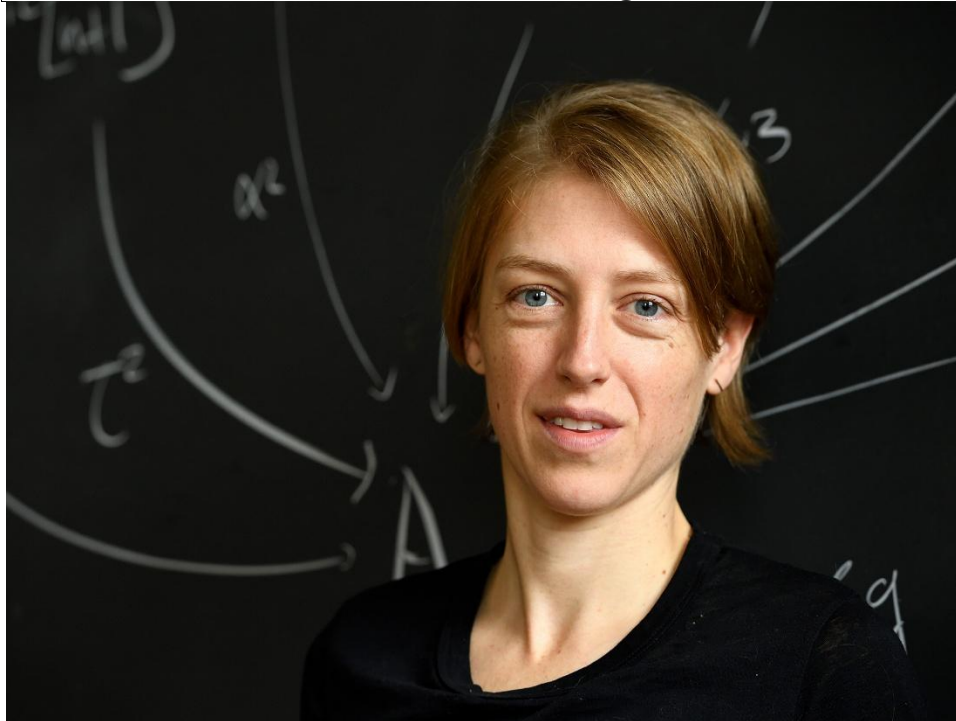
Some mathematicians see theorem provers as a potentially game-changing tool for training undergraduates in proof writing. Others say that getting computers to write proofs is unnecessary for advancing mathematics and probably impossible. But a system that can predict a useful conjecture and prove a new theorem will achieve something new — some machine version of understanding, Szegedy said. And that suggests the possibility of automating reason itself.

Useful Machines

Mathematicians, logicians and philosophers have long argued over what part of creating proofs is fundamentally human, and debates about mechanized

mathematics continue today, especially in the deep valleys connecting computer science and pure mathematics.

For computer scientists, theorem provers are not controversial. They offer a rigorous way to verify that a program works, and arguments about intuition and creativity are less important than finding an efficient way to solve a problem. At the Massachusetts Institute of Technology, for example, the computer scientist [Adam Chlipala](#) has designed theorem-proving tools that `generate()` cryptographic algorithms — traditionally written by humans — to safeguard internet transactions. Already, his group's code is used for the majority of the communication on Google's Chrome browser.



Emily Riehl of Johns Hopkins University uses theorem provers in teaching students and proof assistants in her own work. “Using a proof assistant has changed the way I think about writing proofs,” she said.

Will Kirk/Johns Hopkins University

“You can take any kind of mathematical argument and code it with one tool, and connect your arguments together to create proofs of security,” Chlipala said.

In math, theorem provers have helped produce complicated, calculation-heavy proofs that otherwise would have occupied hundreds of years of mathematicians' lives. The [Kepler conjecture](#), which describes the best way to

stack spheres (or, historically, oranges or cannonballs), offers a telling example. In 1998, [Thomas Hales\(\)](#), together with his student Sam Ferguson, [completed a proof\(\)](#) using a variety of computerized math techniques. The result was so cumbersome — the results took up 3 gigabytes — that 12 mathematicians analyzed it for years before announcing they were 99% certain it was correct.

The Kepler conjecture isn't the only famous question to be solved by machines. [The four-color theorem](#), which says you only need four hues to color any two-dimensional map so that no two adjoining regions share a color, was settled in 1977 by mathematicians using a computer program that churned through five-colored maps to show they could all be reduced to four. And in 2016, a trio of mathematicians used a computer program to prove a longstanding open challenge called the Boolean Pythagorean triples problem, but the initial version of the proof was 200 terabytes in size. With a high-speed internet connection, a person could download it in a little over three weeks.

Complicated Feelings

These examples are often trumpeted as successes, but they've also added to the debate. The computer code proving the four-color theorem, which was settled more than 40 years ago, was impossible for humans to check on their own. "Mathematicians have been arguing ever since whether or not it's a proof," said the mathematician [Michael Harris](#) of Columbia University.

Many mathematicians, like Columbia University's Michael Harris, disagree with the idea that computerized theorem provers are necessary — or that they'll make human mathematicians obsolete.

Béatrice Antolin

Another gripe is that if they want to use theorem provers, mathematicians must first learn to code and then figure out how to express their problem in computer-friendly language — activities that detract from the act of doing math. "By the time I've reframed my question into a form that could fit into this technology, I would have solved the problem myself," Harris said.

Many just don't see a need for theorem solvers in their work. "They have a system, and it's pencil and paper, and it works," said [Kevin Buzzard](#), a mathematician at Imperial College London who three years ago pivoted his work from pure math to focus on theorem provers and formal proofs.

“Computers have done amazing calculations for us, but they have never solved a hard problem on their own,” he said. “Until they do, mathematicians aren’t going to be buying into this stuff.”

But Buzzard and others think maybe they should. For one thing, “computer proofs may not be as alien as we think,” DeDeo said. Recently, together with Scott Viteri, a computer scientist now at Stanford University, he reverse-engineered a handful of famous canonical proofs (including one from Euclid’s *Elements*) and dozens of machine-generated proofs, written using a theorem prover called Coq, to look for commonalities. [They found](#) that the networked structure of machine proofs was remarkably similar to the structure of proofs made by people. That shared trait, he said, may help researchers find a way to get proof assistants to, in some sense, explain themselves.

“Machine proofs may not be as mysterious as they appear,” DeDeo said.

Others say theorem provers can be useful teaching tools, in both computer science and mathematics. At Johns Hopkins University, the mathematician [Emily Riehl](#)() has developed courses in which students write proofs using a theorem prover. “It forces you to be very organized and think clearly,” she said. “Students who write proofs for the first time can have trouble knowing what they need and understanding the logical structure.”

By the time I’ve reframed my question into a form that could fit into this technology, I would have solved the problem myself.

[Michael Harris, Columbia University](#)

Riehl also says that she’s been increasingly using theorem provers in her own work. “It’s not necessarily something you have to use all the time, and will never substitute for scribbling on a piece of paper,” she said, “but using a proof assistant has changed the way I think about writing proofs.”

Theorem provers also offer a way to keep the field honest. In 1999, the Russian American mathematician Vladimir Voevodsky [discovered](#) an error in one of his proofs. From then until his death in 2017, he was a vocal proponent of using computers to check proofs. Hales said that he and Ferguson found hundreds of errors in their original proof when they checked it with computers. Even the very first proposition in Euclid’s *Elements* isn’t perfect. If a machine can help mathematicians avoid such mistakes, why not take advantage of it? (The practical objection, justified or not, is the one suggested by Harris: If mathematicians have to spend their time formalizing math to be understood by a computer, that’s time they’re not spending doing new math.)

But [Timothy Gowers](#), a mathematician and Fields medalist at the University of Cambridge, wants to go even further: He envisions a future in which theorem provers replace human referees at major journals. “I can see it becoming standard practice that if you want your paper to be accepted, you have to get it past an automatic checker,” he said.

Talking to Computers

But before computers can universally check or even devise proofs, researchers first have to clear a significant hurdle: the communication barrier between the language of humans and the language of computers.

Today’s theorem provers weren’t designed to be mathematician-friendly. ATPs, the first type, are generally used to check if a statement is correct, often by testing possible cases. Ask an ATP to verify that a person can drive from Miami to Seattle, for example, and it might search all cities connected by roads leading away from Miami and eventually finding a city with a road leading into Seattle.

Not every mathematician hates theorem provers. Timothy Gowers, of the University of Cambridge, thinks they may one day replace human reviewers at mathematical journals.

The Abel Prize

With an ATP, a programmer can code in all the rules, or axioms, and then ask if a particular conjecture follows those rules. The computer then does all the work. “You just type in the conjecture you want to prove, and you hope you get an answer,” said Daniel Huang, a computer scientist who recently left the University of California, Berkeley, to work at a startup.

But here’s the rub: What an ATP doesn’t do is explain its work. All that calculating happens within the machine, and to human eyes it would look like a long string of 0s and 1s. Huang said it’s impossible to scan the proof and follow the reasoning, because it looks like a pile of random data. “No human will ever look at that proof and be able to say, ‘I get it,’” he said.

ITPs, the second category, have vast data sets containing up to tens of thousands of theorems and proofs, which they can scan to verify that a proof is accurate. Unlike ATPs, which operate in a kind of black box and just spit out an

answer, ITPs require human interaction and even guidance along the way, so they're not as inaccessible. "A human could sit down and understand what the proof-level techniques are," said Huang. (These are the kinds of machine proofs DeDeo and Viteri studied.)

ITPs have become increasingly popular in recent years. In 2017, the trio behind the Boolean Pythagorean triples problem used Coq, an ITP, to create and verify a formal version of their proof; in 2005 Georges Gonthier at Microsoft Research Cambridge used Coq to formalize the four-color theorem. Hales also used ITPs called HOL Light and Isabelle on the formal proof of the Kepler conjecture. ("HOL" stands for "higher-order logic.")

Efforts at the forefront of the field today aim to blend learning with reasoning. They often combine ATPs with ITPs and also integrate machine learning tools to improve the efficiency of both. They envision ATP/ITP programs that can use deductive reasoning — and even communicate mathematical ideas — the same way people do, or at least in similar ways.

The Limits of Reason

Josef Urban thinks that the marriage of deductive and inductive reasoning required for proofs can be achieved through this kind of combined approach. His group has [built theorem provers\(\)](#) guided by [machine learning tools\(\)](#), which allow computers to learn on their own through experience. Over the last few years, they've explored the use of [neural networks](#) — layers of computations that help machines process information through a rough approximation of our brain's neuronal activity. In July, his group reported on [new conjectures\(\)](#) generated by a neural network trained on theorem-proving data.

Urban was partially inspired by Andrej Karpathy, who a few years ago [trained\(\)](#) a neural network to generate mathematical-looking nonsense that looked legitimate to nonexperts. Urban didn't want nonsense, though — he and his group instead designed their own tool to find new proofs after training on millions of theorems. Then they used the network to generate new conjectures and checked the validity of those conjectures using an ATP called E.

The network proposed more than 50,000 new formulas, though tens of thousands were duplicates. "It seems that we are not yet capable of proving the more interesting conjectures," Urban said.

[Machines] will learn how to do their own prompts.

Timothy Gowers, University of Cambridge

Szegedy at Google Research sees the challenge of automating reasoning in computer proofs as a subset of a much bigger field: natural language processing, which involves pattern recognition in the usage of words and sentences. (Pattern recognition is also the driving idea behind computer vision, the object of Szegedy's previous project at Google.) Like other groups, his team wants theorem provers that can find and explain useful proofs.

Inspired by the rapid development of AI tools like [AlphaZero](#) — the DeepMind program that can defeat humans at chess, Go and shogi — Szegedy's group wants to capitalize on recent advances in language recognition to write proofs. Language models, he said, can demonstrate surprisingly solid mathematical reasoning.

His group at Google Research recently [described\(\)](#) a way to use language models — which often use neural networks — to generate new proofs. After training the model to recognize a kind of treelike structure in theorems that are known to be true, they ran a kind of free-form experiment, simply asking the network to generate and prove a theorem without any further guidance. Of the thousands of generated conjectures, about 13% were both provable and new (meaning they didn't duplicate other theorems in the database). The experiment, he said, suggests that the neural net could teach itself a kind of understanding of what a proof looks like.

"Neural networks are able to develop an artificial style of intuition," Szegedy said.

Of course, it's still unclear whether these efforts will fulfill Cohen's prophecy from over 40 years ago. Gowers has said that he thinks computers will be able to out-reason mathematicians by 2099. At first, he predicts, mathematicians will enjoy a kind of golden age, "when mathematicians do all the fun parts and computers do all the boring parts. But I think it will last a very short time."

After all, if the machines continue to improve, and they have access to vast amounts of data, they should become very good at doing the fun parts, too. "They will learn how to do their own prompts," Gowers said.

Harris disagrees. He doesn't think computer provers are necessary, or that they will inevitably "make human mathematicians obsolete." If computer scientists are ever able to program a kind of synthetic intuition, he says, it still won't rival that of humans. "Even if computers understand, they don't understand in a human way."

ARTIFICIAL INTELLIGENCE

Common Sense Comes Closer to Computers

The problem of common-sense reasoning has plagued the field of artificial intelligence for over 50 years. Now a new approach, borrowing from two disparate lines of thinking, has made important progress.

Add a match to a pile of wood. What do you get? For a human, it's easy. But machines have long lacked the common-sense reasoning abilities needed to figure it out.

[Guillem Casasús Xercavins\(\)](#) for Quanta Magazine

Introduction

One evening last October, the artificial intelligence researcher [Gary Marcus\(\)](#) was amusing himself on his iPhone by making a state-of-the-art neural network look stupid. Marcus' target, a deep learning network called GPT-2, had recently become famous for its uncanny ability to generate plausible-sounding English prose with just a sentence or two of prompting. When [journalists at The Guardian fed it text from a report on Brexit\(\)](#), GPT-2 wrote entire newspaper-style paragraphs, complete with convincing political and geographic references. Marcus, a prominent critic of AI hype, gave the neural network a pop quiz. He typed the following [into GPT-2](#):

What happens when you stack kindling and logs in a fireplace and then drop some matches is that you typically start a ...

Surely a system smart enough to [contribute to The New Yorker](#)) would have no trouble completing the sentence with the obvious word, “fire.” GPT-2 responded with “ick.” In another attempt, it suggested that dropping matches on logs in a fireplace would start an “irc channel full of people.” Marcus wasn’t surprised. Common-sense reasoning — the ability to make mundane inferences using basic knowledge about the world, like the fact that “matches” plus “logs” usually equals “fire” — has resisted AI researchers’ efforts for decades. [Marcus posted the exchanges to his Twitter account](#) with his own added commentary: “LMAO,” internet slang for a derisive chortle. Neural networks might be impressive linguistic mimics, but they clearly lacked basic common sense. Minutes later, [Yejin Choi](#) saw Marcus’ snarky tweet. The timing was awkward. Within the hour Choi was scheduled to give a talk at a prominent AI conference on her latest research project: a system, nicknamed COMET, that was designed to use an earlier version of GPT-2 to perform common-sense reasoning.

Quickly, Choi — a computer scientist at the University of Washington and the Allen Institute for Artificial Intelligence — fed COMET the same prompt Marcus had used (with its wording slightly modified to match COMET’s input format):

Gary stacks kindling and logs and drops some matches

COMET generated 10 inferences about why Gary might be dropping the matches. Not all of the responses made sense, but the first two did: He “wanted to start a fire” or “to make a fire.” Choi [tweeted the results in reply to Marcus\(\)](#) and strode up to the podium to include them in her presentation. “It seemed only appropriate,” she said.

Two Paths to Common Sense

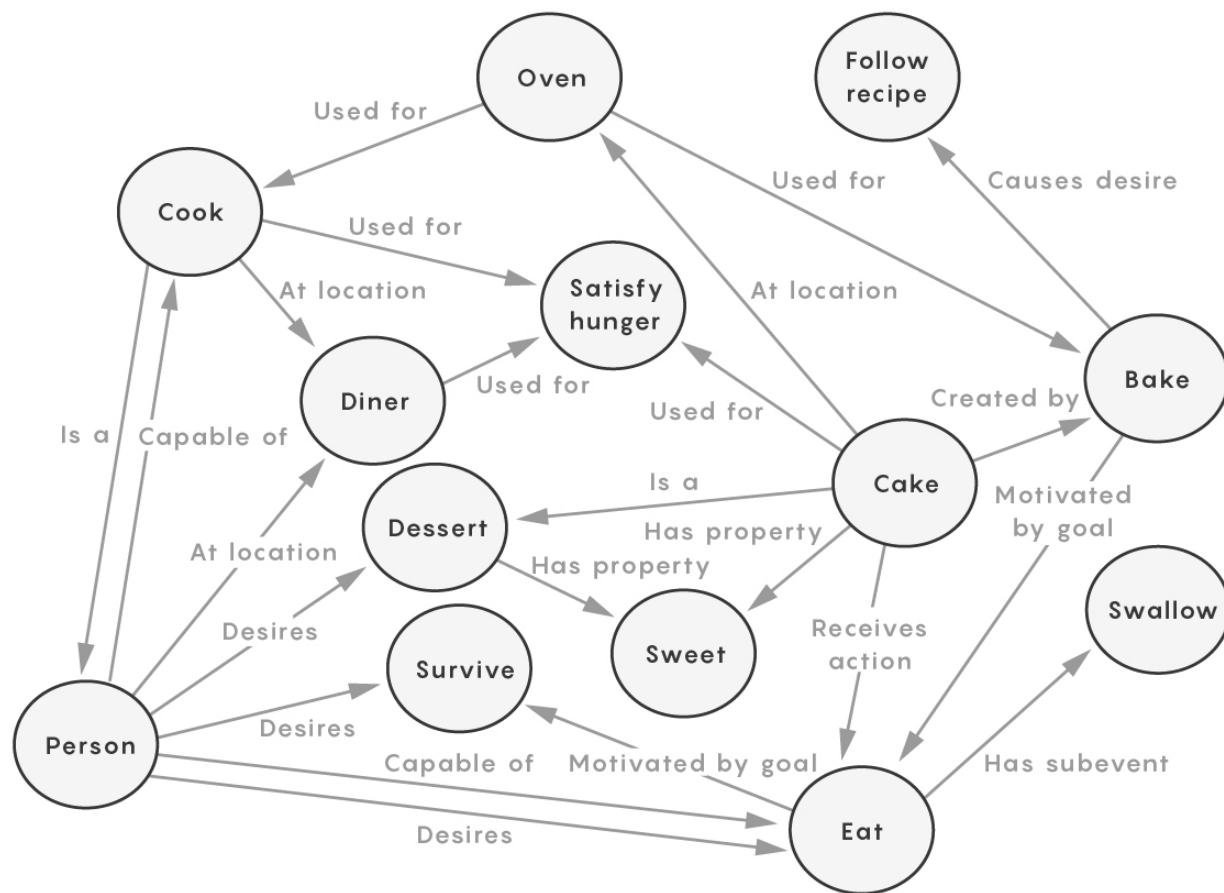
Common sense has been called the [“dark matter of AI\(”](#) — both essential and frustratingly elusive. That’s because common sense consists of implicit information — the broad (and broadly shared) set of unwritten assumptions and rules of thumb that humans automatically use to make sense of the world. For example, consider the following scenario:

A man went to a restaurant. He ordered a steak. He left a big tip.

If you were asked what he ate, the answer — steak — comes effortlessly. But nowhere in that little scene is it ever stated that the man actually ate anything. When [Ray Mooney\(\)](#), director of the Artificial Intelligence Laboratory at the University of Texas, Austin, pointed this out after giving me the same pop quiz, I didn't believe him at first. "People don't even realize that they're doing this," he said. Common sense lets us read between the lines; we don't need to be explicitly told that food is typically eaten in restaurants after people order and before they leave a tip.

Computers do. It's no wonder that common-sense reasoning emerged as a primary concern of AI research in 1958 (in a paper titled "[Programs With Common Sense\(\)](#)"), not long after the field of AI was born. "In general, you can't do natural language understanding or vision or planning without it," said [Ernest Davis\(\)](#), a computer scientist at New York University who has studied common sense in AI since the 1980s.

Still, progress has been infamously slow. At first, researchers tried to translate common sense into the language of computers: logic. They surmised that if all the unwritten rules of human common sense could be written down, computers should be able to use them to reason with in the same way that they do arithmetic. This symbolic approach, which came to be known as "good old-fashioned artificial intelligence" (or GOFAI), enabled some early successes, but its handcrafted approach didn't scale. "The amount of knowledge which can be conveniently represented in the formalisms of logic is kind of limited in principle," said [Michael Witbrock\(\)](#), an AI researcher at the University of Auckland in New Zealand. "It turned out to be a truly overwhelming task."



Even modest attempts to map all possible logical relationships quickly run into trouble. Some of the relationships above always hold (for example, swallowing is always a part of eating). Some hold only occasionally (a person eats at a diner). Some are inconsistent (a person can not eat a cake while it is also in the oven). And nodes such as “cook” can mean both a person who cooks and the activity of cooking.

doi: [10.1109/MIS.2009.72](https://doi.org/10.1109/MIS.2009.72)()

Deep learning with neural networks seemed to offer an alternative. These AI systems, designed to mimic the interconnected layers of neurons in biological brains, learn patterns without requiring programmers to specify them in advance. Over the past decade, increasingly sophisticated neural networks, trained with copious amounts of data, have revolutionized computer vision and natural language processing. But for all their flexibility and apparent intellectual power — neural networks can now steer cars in highway traffic and beat world-class players at chess and Go — these systems remain notorious for their own silly (and occasionally [fatal\(\)](#)) lapses in ordinary common sense. “Acquiring it, representing it, reasoning with it — it’s all hard,” Davis said.

Now, Choi and her collaborators have united these approaches. [COMET\(\)](#) (short for “commonsense transformers”) extends GOF AI-style symbolic reasoning with the latest advances in neural language

modeling — a kind of deep learning that aims to imbue computers with a [statistical “understanding” of written language](#). COMET works by reimagining common-sense reasoning as a process of generating plausible (if imperfect) responses to novel input, rather than making airtight deductions by consulting a vast encyclopedia-like database.

“It tries to blend two very fundamentally different approaches to AI,” said Mooney, who is already using COMET in his own research. “It’s an interesting new direction that says, ‘Hey, there’s a middle road there.’” [Leora Morgenstern\(\)](#), an expert in common-sense reasoning and AI at the Palo Alto Research Center who has spent decades researching symbolic approaches to the problem, thinks that the ideas behind COMET can help move the field forward. “One of the reasons I’m so excited about what Yejin is doing is I think it will inject new life into the common-sense reasoning community,” she said. “Deep learning is really, really powerful — let’s figure out how to harness it for common sense.”

Endless Unwritten Rules

Common sense is easier to detect than to define. According to Witbrock, the phrase “common sense” can mean both a kind of knowledge and an attitude toward that knowledge. “I would say [it’s] broadly reusable background knowledge that’s not specific to a particular subject area,” he said. “It’s knowledge that you ought to have.” Like, for example, the fact that people eat food in restaurants, rather than just ordering and paying for it; or that dropping matches on a pile of stacked logs implies that one is trying to light a fire.

The implicit nature of most common-sense knowledge makes it difficult and tedious to represent explicitly. “What you learn when you’re two or four years old, you don’t really ever put down in a book,” said Morgenstern. Nevertheless, early AI researchers believed that bridging this gap was possible. “It was like, ‘Let’s write down all the facts about the world. Surely there’s only a couple million of them,’” said [Ellie Pavlick\(\)](#), a computer scientist at Brown University. Constructing such a resource, known as a knowledge base, has traditionally been the first step in any approach to automating common-sense reasoning.

What you learn when you’re two or four years old, you don’t really ever put down in a book.

[Leora Morgenstern](#)

Building up a sufficient number of obvious facts is harder than it sounds. A common-sense reasoning project called Cyc began in 1984 with the modest-sounding goal of encoding the implicit common-sense knowledge necessary to represent 400 encyclopedia articles. It [never stopped\(\)](#). More than three decades later, Cyc's knowledge base — encoded in a dense, custom-designed logical notation — [contains\(\)](#) “millions of collections and concepts, and more than 25 million assertions.” Yet a [2015 review article\(\)](#) by Davis and Marcus stated that “Cyc has had comparatively little impact on AI research.” Subsequent attempts to write entries for a knowledge base — or to create one by mining documents using machine learning — have failed to crack the common-sense reasoning problem.

Why? For one thing, “there’s always exceptions to every case,” Pavlick explained. “If I hear some statement like ‘it’s raining,’ I could infer that if I go outside I’ll get wet, but not if [I’m] underneath something.” Other exceptions are harder to anticipate. A knowledge base like Cyc may contain dozens of statements about what typically happens when a person orders food in a restaurant. But what about the potentially endless list of infrequent or unusual things that could happen in that scenario, like leaving without paying the check, or starting a food fight? “Coverage is never-ending,” said Choi. “Therefore, purely symbolic knowledge-based approaches are entirely doomed.”

Even if it were possible to build a knowledge base 100 or 1,000 times as comprehensive as any previous attempt, the system would still suffer from another intellectual shortcoming: the so-called brittleness problem. That’s because common sense, like natural language, remains fundamentally fuzzy. When a server asks a diner, “Are you still working on that?” we understand them to mean “Are you still eating what’s on your plate?” But if the server asks the same question to a line cook preparing an overdue order, it means something else entirely. So is a restaurant a place where people “work” on things? Are “eating” and “working” distinct concepts?

It all depends. That’s the brittleness problem: Sharply defined relations within a knowledge base may enable powerful, reliable reasoning abilities, as long as those conceptual edges are respected. But these symbolic systems, no matter how varied and rich, inevitably fail to capture the natural ambiguities and associative overlaps that often occur in human common-sense reasoning. “To the extent that we [use] symbols,” Pavlick said, “we’re quite fluid with them.”

Going Neural

Choi didn't start working on common sense because she wanted to tilt at windmills. When she joined the Allen Institute in 2018, she "had a hunch" that neural networks could enable new progress where knowledge bases had stalled on their own. She just didn't know exactly how. She didn't want to write off previous symbolic approaches completely, either. "All the past research was based on a lack of data," she said, or a lack of computing resources. "So I figured I'd just withhold my judgment until I properly tried different routes."

With an open mind, Choi and her colleagues began to assemble their own knowledge base called [Atomic\(\)](#) (short for "atlas of machine commonsense"). "Basically, I wanted to write a textbook for neural networks to learn faster about the world," Choi said. "Then things happened simultaneously — as we had this knowledge [base] built, GPT-2 came out."

Yejin Choi has begun to add a visual component to common-sense reasoning.
Allen Institute for AI

That neural network, released in February 2019, was just one in a wave of "pre-trained language models" that began to revolutionize how computers process natural language. These systems don't contain neatly organized linguistic symbols or rules. Instead, they statistically smear their representations of language across millions or billions of parameters within a neural network. This property makes such systems difficult to interpret, but it also makes them robust: They can generate predictions based on noisy or ambiguous input without breaking. When fine-tuned to perform a specific task — like answering written questions or paraphrasing text — language models [even appear to understand at least some of what they're reading](#).

Choi now saw a way to put her hunch about neural networks and common sense into action.

What would happen if a language model were given additional training using a common-sense knowledge base, like Atomic? Could the neural network learn to fill in Atomic's gaps with plausible commonsense inferences all on its own, just as GPT-2 learned how to automatically generate plausible news articles? "It's almost weird that nobody tried this before," Choi said. "It's almost as if nobody bothered because they were so sure this would never work."

When Choi (and her collaborators Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya and Asli Celikyilmaz) fine-tuned a neural language model with the common-sense knowledge encoded in Atomic, they created COMET. Its fusion of symbolic reasoning with a neural network tries to solve the coverage and brittleness problems at the same time. Anyone can `type a prompt into COMET()` in everyday language. If the event is already represented in the system's `common-sense knowledge base()` (like the fact that ordering food in a restaurant usually involves eating it), COMET can simply reason with that preexisting information. For everything else, the neural language model makes its best guess.

Those guesses are surprisingly good. On average, 77.5% of the novel responses generated by COMET — that is, inferences that come from the neural network, rather than from the preexisting knowledge base — were deemed “plausible” by teams of human evaluators. That’s less than 10 percentage points shy of human-level performance. (Evaluators found 86% of knowledge-base entries written by humans to be plausible.) When COMET was given the prompt “PersonX gives PersonY some pills,” it guessed that PersonX wanted to help; when it was told that “PersonX murders PersonY’s wife,” COMET suggested that PersonX wanted to hide the body.

These examples showed how COMET could handle input beyond the limits of its built-in common-sense “coverage.” But what about the brittleness problem? While interviewing Choi late last year at her lab in Seattle, I gave COMET a prompt phrased in my 5-year-old daughter’s patois: “Daddy goed to work.”

Choi frowned. “That may be tricky,” she said. But COMET took it in stride, suggesting that “Daddy” wanted to “make money,” “do their job” and “get a paycheck”; that he is seen as “hardworking,” “motivated” and “dutiful”; and that as a result, others feel “proud,” “grateful” and — in an amusingly plausible response, given that the request was written in kindergartner-speak — “annoyed.” (My daughter has certainly expressed that sentiment when I leave for work instead of playing with her.) “This wouldn’t work with Cyc, for sure,” Choi remarked. “Unless someone hand-codes that ‘goed’ means ‘went’ — which we never did.”

Ladders and Rockets

There's a quip Gary Marcus `likes()` to `use()` to put progress in AI into context: "Just because you can build a better ladder doesn't mean you can build a ladder to the moon." To him and others, COMET's approach suffers from a fundamental limitation of deep learning: "`statistics ≠ understanding()`." "You can see that [COMET] does a decent job of guessing some of the parameters of what a sentence might entail, but it doesn't do so in a consistent way," Marcus wrote via email. Just as no ladder, no matter how tall, can ever hope to reach the moon, no neural network — no matter how deft at mimicking language patterns — ever really "knows" that dropping lit matches on logs will typically start a fire.

Choi, surprisingly, agrees. She acknowledged that COMET "relies on surface patterns" in its training data, rather than actual understanding of concepts, to generate its responses. "But the fact that it's really good at surface patterns is a good thing," she said. "It's just that we've got to supply it with more informative surface patterns."

It was like, 'Let's write down all the facts about the world. Surely there's only a couple million of them.'

Ellie Pavlick

What might those more informative patterns look like? Some researchers argue that in order to build real common sense into computers, we will need to make use of phenomena outside language itself, like visual perceptions or embodied sensations. These more direct first-person representations may be the foundation of common sense, with language acting as a secondary layer.

"If I lived in a world where there were no other people [to talk to], I could still have common sense — I'd still understand how the world works and have expectations over what I should see and shouldn't see," said Pavlick, who is currently studying how to teach AI systems common sense `by interacting with them in virtual reality()`. To her, COMET represents "really exciting progress, but what's missing is the actual reference aspect. The word 'apple' is not an apple. That meaning has to exist in some form that's not the language itself." Nazneen Rajani(), a senior research scientist at Salesforce, is pursuing a similar goal, but she believes the full potential of neural language models is far from tapped. She's investigating whether they can learn to reason about common-sense scenarios involving basic physics, like the fact that tipping

over a jar with a ball inside will typically cause the ball to drop out. “The real world is really complicated,” Rajani said. “But natural language is like a low-dimensional proxy for how the real world works.” Sure, neural networks can be taught to predict the next word from a text prompt, but that shouldn’t be their limit. “They can learn more complex stuff.”

Choi and her colleagues are also working on ways to [augment COMET\(\)](#) with [labeled visual scenes\(\)](#) instead of just text. “We took all these images from movies or TV shows where some interesting things are happening,” Choi said. “The annotations look great; the model predictions look exciting.”

I asked Choi if COMET’s approach — combining incrementally better neural networks with improved common-sense knowledge bases — was still, essentially, building a ladder to the moon. She conceded that her dream would be to have a neural network that could learn from knowledge bases without human supervision, the same way language models like GPT-2 already learn by ingesting reams of raw text.

But just as Winston Churchill quipped that “democracy is the worst form of government except for all those other forms that have been tried,” Choi considers COMET’s flawed but promising approach to be “a fair deal.” Even if these neural networks can’t reach the stars, she thinks they’re the only way to get off the ground. “Without that, we are not going anywhere,” she said. “With [knowledge bases] alone, we cannot do anything. It’s COMET that can actually fly in the air.”

The Physics Principle That Inspired Modern AI Art

Diffusion models generate incredible images by learning to reverse the process that, among other things, causes ink to spread through water.

Introduction

Ask DALL·E 2, an image generation system created by OpenAI, to paint a picture of “goldfish slurping Coca-Cola on a beach,” and it will spit out surreal images of exactly that. The program would have encountered images of beaches, goldfish and Coca-Cola during training, but it’s highly unlikely it would have seen one in which all three came together. Yet DALL·E 2 can assemble the concepts into something that might have made Dalí proud.

DALL·E 2 is a type of generative model — a system that attempts to use training data to generate something new that’s comparable to the data in terms of quality and variety. This is one of the hardest problems in machine learning, and getting to this point has been a difficult journey.

The first important generative models for images used an approach to artificial intelligence called a neural network — a program composed of many layers of computational units called artificial neurons. But even as the quality of their images got better, the models proved unreliable and hard to train. Meanwhile, a powerful generative model — created by a postdoctoral researcher with a passion for physics — lay dormant, until two graduate students made technical breakthroughs that brought the beast to life.

DALL·E 2 is such a beast. The key insight that makes DALL·E 2’s images possible — as well as those of its competitors Stable Diffusion and Imagen — comes from the world of physics. The system that underpins them, known as a diffusion model, is heavily inspired by nonequilibrium thermodynamics, which governs phenomena like the spread of fluids and gases. “There are a lot of techniques that were initially invented by physicists and now are very important in machine learning,” said [Yang Song\(\)](#), a machine learning researcher at OpenAI.

The power of these models has rocked industry and users alike. “This is an exciting time for generative models,” said [Anima Anandkumar](#), a computer scientist at the California Institute of Technology and senior director of machine learning research at Nvidia. And while the realistic-looking images created by diffusion models can sometimes perpetuate social and cultural biases, she said, “we have demonstrated that generative models are useful for downstream tasks [that] improve the fairness of predictive AI models.”

High Probabilities

To understand how creating data works for images, let’s start with a simple image made of just two adjacent grayscale pixels. We can fully describe this image with two values, based on each pixel’s shade (from zero being completely black to 255 being completely white). You can use these two values to plot the image as a point in 2D space.

If we plot multiple images as points, clusters may emerge — certain images and their corresponding pixel values that occur more frequently than others. Now imagine a surface above the plane, where the height of the surface corresponds to how dense the clusters are. This surface maps out a probability. You’re most likely to find individual data points underneath the highest part of the surface, and few where the surface is lowest.\

DALL·E 2 produced these images of “goldfish slurping Coca-Cola on a beach.” The program, created by OpenAI, had likely never encountered similar images, but could still generate them on its own.

DALL·E 2

Now you can use this probability distribution to generate new images. All you need to do is randomly generate new data points while adhering to the restriction that you generate more probable data more often — a process called “sampling” the distribution. Each new point is a new image.

The same analysis holds for more realistic grayscale photographs with, say, a million pixels each. Only now, plotting each image requires not two axes, but a million. The probability distribution over such images will be some complex million-plus-one-dimensional surface. If you sample that distribution, you’ll produce a million pixel values. Print those pixels on a sheet of paper, and the image will likely look like a photo from the original data set.

The challenge of generative modeling is to learn this complicated probability distribution for some set of images that constitute training data. The distribution is useful partly because it captures extensive information about the data, and partly because researchers can combine probability distributions over different types of data (such as text and images) to compose surreal outputs, such as a goldfish slurping Coca-Cola on a beach. “You can mix and match different concepts ... to create entirely new scenarios that were never seen in training data,” said Anandkumar.

In 2014, a model called a generative adversarial network (GAN) became the first to produce realistic images. “There was so much excitement,” said Anandkumar. But GANs are hard to train: They may not learn the full probability distribution and can get locked into producing images from only a subset of the distribution. For example, a GAN trained on images of a variety of animals may generate only pictures of dogs.

Machine learning needed a more robust model. [Jascha Sohl-Dickstein\(\)](#), whose work was inspired by physics, would provide one.

Blobs of Excitement

Around the time GANs were invented, Sohl-Dickstein was a postdoc at Stanford University working on generative models, with a side interest in nonequilibrium thermodynamics. This branch of physics studies systems not in thermal equilibrium — those that exchange matter and energy internally and with their environment.

An illustrative example is a drop of blue ink diffusing through a container of water. At first, it forms a dark blob in one spot. At this point, if you want to calculate the probability of finding a molecule of ink in some small volume of the container, you need a probability distribution that cleanly models the initial state, before the ink begins spreading. But this distribution is complex and thus hard to sample from.

Eventually, however, the ink diffuses throughout the water, making it pale blue. This leads to a much simpler, more uniform probability distribution of molecules that can be described with a straightforward mathematical expression. Nonequilibrium thermodynamics describes the probability distribution at each step in the diffusion process. Crucially, each step is

reversible — with small enough steps, you can go from a simple distribution back to a complex one.

Jascha Sohl-Dickstein created a new approach for generative modeling based on the principles of diffusion.

Asako Miyakawa

Sohl-Dickstein used the principles of diffusion to develop an algorithm for generative modeling. The idea is simple: The algorithm first turns complex images in the training data set into simple noise — akin to going from a blob of ink to diffuse light blue water — and then teaches the system how to reverse the process, turning noise into images.

Here's how it works. First, the algorithm takes an image from the training set. As before, let's say that each of the million pixels has some value, and we can plot the image as a dot in million-dimensional space. The algorithm adds some noise to each pixel at every time step, equivalent to the diffusion of ink after one small time step. As this process continues, the values of the pixels bear less of a relationship to their values in the original image, and the pixels look more like a simple noise distribution. (The algorithm also nudges each pixel value a smidgen toward the origin, the zero value on all those axes, at each time step. This nudge prevents pixel values from growing too large for computers to easily work with.)

Do this for all images in the data set, and an initial complex distribution of dots in million-dimensional space (which cannot be described and sampled from easily) turns into a simple, normal distribution of dots around the origin.

“The sequence of transformations very slowly turns your data distribution into just a big noise ball,” said Sohl-Dickstein. This “forward process” leaves you with a distribution you can sample from with ease.

Yang Song helped came up with a novel technique to generate images by training a network to effectively unscramble noisy images.

Courtesy of Yang Song

Next is the machine learning part: Give a neural network the noisy images obtained from a forward pass and train it to predict the less noisy images that came one step earlier. It'll make mistakes at first, so you tweak the parameters

of the network so it does better. Eventually, the neural network can reliably turn a noisy image, which is representative of a sample from the simple distribution, all the way into an image representative of a sample from the complex distribution.

The trained network is a full-blown generative model. Now you don't even need an original image on which to do a forward pass: You have a full mathematical description of the simple distribution, so you can sample from it directly. The neural network can turn this sample — essentially just static — into a final image that resembles an image in the training data set.

Sohl-Dickstein recalls the first outputs of his diffusion model. “You’d squint and be like, ‘I think that colored blob looks like a truck,’” he said. “I’d spent so many months of my life staring at different patterns of pixels and trying to see structure that I was like, ‘This is way more structured than I’d ever gotten before.’ I was very excited.”

Envisioning the Future

Sohl-Dickstein published his [diffusion model algorithm\(\)](#) in 2015, but it was still far behind what GANs could do. While diffusion models could sample over the entire distribution and never get stuck spitting out only a subset of images, the images looked worse, and the process was much too slow. “I don’t think at the time this was seen as exciting,” said Sohl-Dickstein.

It would take two students, neither of whom knew Sohl-Dickstein or each other, to connect the dots from this initial work to modern day diffusion models like DALL·E 2. The first was Song, a doctoral student at Stanford at the time. In 2019, he and his adviser [published a novel method\(\)](#) for building generative models that didn’t estimate the probability distribution of the data (the high-dimensional surface). Instead, it estimated the gradient of the distribution (think of it as the slope of the high-dimensional surface).

ARTIFICIAL INTELLIGENCE

Self-Taught AI Shows Similarities to How the Brain Works

Song found his technique worked best if he first perturbed each image in the training data set with increasing levels of noise, then asked his neural network to predict the original image using gradients of the distribution, effectively denoising it. Once trained, his neural network could take a noisy image sampled from a simple distribution and progressively turn that back into an image representative of the training data set. The image quality was great, but his machine learning model was painfully slow to sample. And he did this with no knowledge of Sohl-Dickstein's work. "I was not aware of diffusion models at all," said Song. "After our 2019 paper was published, I received an email from Jascha. He pointed out to me that [our models] have very strong connections."

In 2020, the second student saw those connections and realized that Song's work could improve Sohl-Dickstein's diffusion models. [Jonathan Ho](#)() had recently finished his doctoral work on generative modeling at the University of California, Berkeley, but he continued working on it. "I thought it was the most mathematically beautiful subdiscipline of machine learning," he said.

Ho redesigned and updated Sohl-Dickstein's diffusion model with some of Song's ideas and other advances from the world of neural networks. "I knew that in order to get the community's attention, I needed to make the model generate great-looking samples," he said. "I was convinced that this was the most important thing I could do at the time."

His intuition was spot on. Ho and his colleagues announced this new and improved diffusion model in 2020, in a paper titled "[Denoising Diffusion](#)

[Probabilistic Models\(\)](#).” It quickly became such a landmark that researchers now refer to it simply as DDPM. According to one benchmark of image quality — which compares the distribution of generated images to the distribution of training images — these models matched or surpassed all competing generative models, including GANs. It wasn’t long before the big players took notice. Now, DALL·E 2, Stable Diffusion, Imagen and other commercial models all use some variation of DDPM.

Jonathan Ho and his colleagues combined Sohl-Dickstein and Song’s methods to make possible modern diffusion models, such as DALL·E 2.

Courtesy of Jonathan Ho

Modern diffusion models have one more key ingredient: large language models (LLMs), such as GPT-3. These are generative models trained on text from the internet to learn probability distributions over words instead of images. In 2021, Ho — now a research scientist at a stealth company — and his colleague [Tim Salimans\(\)](#) at Google Research, along with other teams elsewhere, showed how to combine information from an LLM and an image-generating diffusion model to use text (say, “goldfish slurping Coca-Cola on a beach”) to guide the process of diffusion and hence image generation. This process of “guided diffusion” is behind the success of text-to-image models, such as DALL·E 2.

“They are way beyond my wildest expectations,” said Ho. “I’m not going to pretend I saw all this coming.”

Generating Problems

As successful as these models have been, images from DALL·E 2 and its ilk are still far from perfect. Large language models can reflect cultural and societal biases, such as racism and sexism, in the text they generate. That’s because they are trained on text taken off the internet, and often such texts contain racist and sexist language. LLMs that learn a probability distribution over such text become imbued with the same biases. Diffusion models are also trained on un-curated images taken off the internet, which can contain similarly biased data. It’s no wonder that combining LLMs with today’s diffusion models can sometimes result in images reflective of society’s ills.

Anandkumar has firsthand experience. When she tried to generate stylized avatars of herself using a diffusion model-based app, she was shocked. “So

[many] of the images were highly sexualized,” she said, “whereas the things that it was presenting to men weren’t.” She’s [not alone\(\)](#).

These biases can be lessened by curating and filtering the data (an extremely difficult task, given the immensity of the data set), or by putting checks on both the input prompts and the outputs of these models. “Of course, nothing is a substitute for carefully and extensively safety-testing” a model, Ho said. “This is an important challenge for the field.”

Despite such concerns, Anandkumar believes in the power of generative modeling. “I really like Richard Feynman’s quote: ‘What I cannot create, I do not understand,’” she said. An increased understanding has enabled her team to develop generative models to produce, for example, synthetic training data of under-represented classes for predictive tasks, such as darker skin tones for facial recognition, helping improve fairness. Generative models may also give us insights into how our brains deal with noisy inputs, or how they conjure up mental imagery and contemplate future action. And building more sophisticated models could endow AIs with similar capabilities.

“I think we are just at the beginning of the possibilities of what we can do with generative AI,” said Anandkumar.

ARTIFICIAL INTELLIGENCE

Foundations Built for a General Theory of Neural Networks

Neural networks can be as unpredictable as they are powerful. Now mathematicians are beginning to reveal how a neural network’s form will influence its function.

Introduction

When we design a skyscraper we expect it will perform to specification: that the tower will support so much weight and be able to withstand an earthquake of a certain strength.

But with one of the most important technologies of the modern world, we’re effectively building blind. We play with different designs, tinker with different

setups, but until we take it out for a test run, we don't really know what it can do or where it will fail.

This technology is the neural network, which underpins today's most advanced artificial intelligence systems. Increasingly, neural networks are moving into the core areas of society: They determine what we learn of the world through our social media feeds, they help [doctors diagnose illnesses](#)(), and they even [influence](#)() whether a person convicted of a crime will spend time in jail. Yet "the best approximation to what we know is that we know almost nothing about how neural networks actually work and what a really insightful theory would be," said [Boris Hanin](#)(), a mathematician at Texas A&M University and a visiting scientist at Facebook AI Research who studies neural networks.

He likens the situation to the development of another revolutionary technology: the steam engine. At first, steam engines weren't good for much more than pumping water. Then they powered trains, which is maybe the level of sophistication neural networks have reached. Then scientists and mathematicians developed a theory of thermodynamics, which let them understand exactly what was going on inside engines of any kind. Eventually, that knowledge took us to the moon.

"First you had great engineering, and you had some great trains, then you needed some theoretical understanding to go to rocket ships," Hanin said.

Within the sprawling community of neural network development, there is a small group of mathematically minded researchers who are trying to build a theory of neural networks — one that would explain how they work and guarantee that if you construct a neural network in a prescribed manner, it will be able to perform certain tasks.

This work is still in its very early stages, but in the last year researchers have produced several papers which elaborate the relationship between form and function in neural networks. The work takes neural networks all the way down to their foundations. It shows that long before you can certify that neural networks can drive cars, you need to prove that they can multiply.

The Best Brain Recipe

Neural networks aim to mimic the human brain — and one way to think about the brain is that it works by accreting smaller abstractions into larger ones. Complexity of thought, in this view, is then measured by the range of smaller abstractions you can draw on, and the number of times you can combine lower-level abstractions into higher-level abstractions — like the way we learn to distinguish dogs from birds.

“For a human, if you’re learning how to recognize a dog you’d learn to recognize four legs, fluffy,” said [Maithra Raghu\(\)](#), a doctoral student in computer science at Cornell University and a member of [Google Brain\(\)](#). “Ideally we’d like our neural networks to do the same kinds of things.”

Maithra Raghu, a member of Google Brain, has been identifying principles that explain how neural networks operate.

Arun Chaganty

Abstraction comes naturally to the human brain. Neural networks have to work for it. As with the brain, neural networks are made of building blocks called “neurons” that are connected in various ways. (The neurons in a neural network are inspired by neurons in the brain but do not imitate them directly.) Each neuron might represent an attribute, or a combination of attributes, that the network considers at each level of abstraction.

When joining these neurons together, engineers have many choices to make. They have to decide how many layers of neurons the network should have (or how “deep” it should be). Consider, for example, a neural network with the task of recognizing objects in images. The image enters the system at the first layer. At the next layer, the network might have neurons that simply detect edges in the image. The next layer combines lines to identify curves in the image. Then the next layer combines curves into shapes and textures, and the final layer processes shapes and textures to reach a conclusion about what it’s looking at: woolly mammoth!

“The idea is that each layer combines several aspects of the previous layer. A circle is curves in many different places, a curve is lines in many different places,” said [David Rolnick\(\)](#), a mathematician at the University of Pennsylvania.

Engineers also have to decide the “width” of each layer, which corresponds to the number of different features the network is considering at each level of

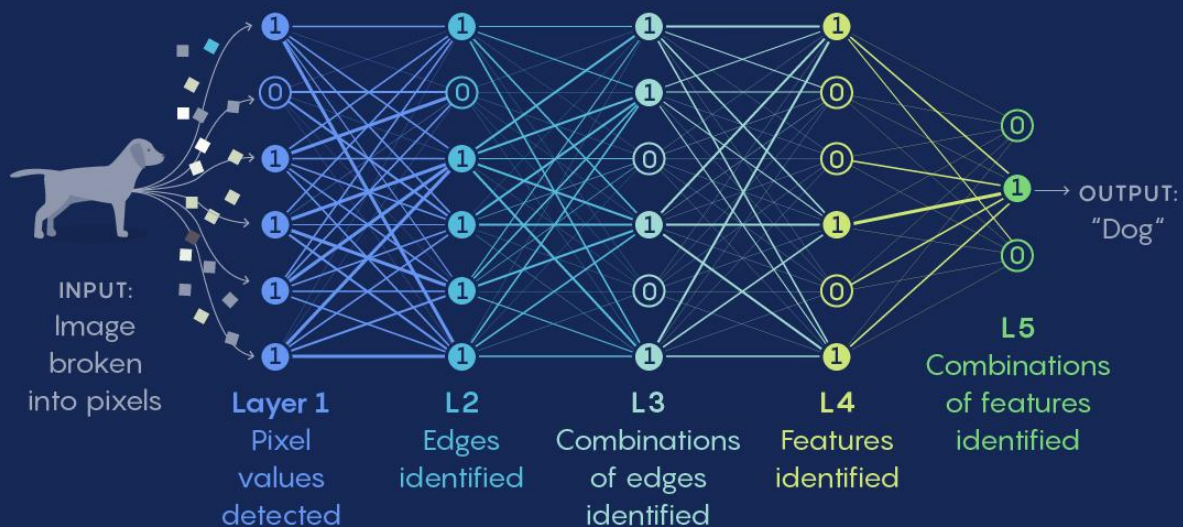
abstraction. In the case of image recognition, the width of the layers would be the number of types of lines, curves or shapes it considers at each level.

Beyond the depth and width of a network, there are also choices about how to connect neurons within layers and between layers, and how much weight to give each connection.

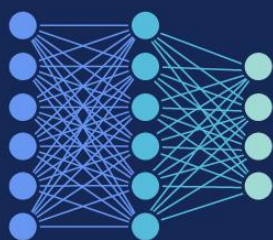
So if you have a specific task in mind, how do you know which neural network architecture will accomplish it best? There are some broad rules of thumb. For image-related tasks, engineers typically use “convolutional” neural networks, which feature the same pattern of connections between layers repeated over and over. For natural language processing — like speech recognition, or language generation — engineers have found that “recurrent” neural networks seem to work best. In these, neurons can be connected to non-adjacent layers.

How to Design a Neural Network

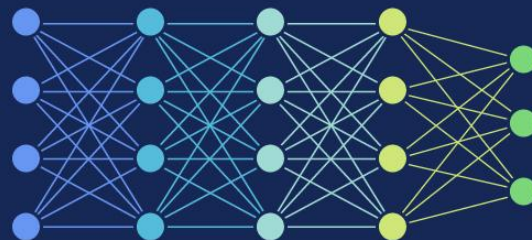
Neural networks pass an input, like an image, through multiple layers of digital neurons. Each layer reveals additional features of the input. Mathematicians are revealing how a network's architecture — how many neurons and layers it has and how they're connected — determines the kinds of tasks that the neural network will be good at.



When data is fed into a network, each artificial neuron that fires (labeled "1") transmits signals to certain neurons in the next layer, which are likely to fire if multiple signals are received. This process reveals abstract information about the input.



A **SHALLOW NETWORK** has few layers but many neurons per layer. These "expressive" networks are computationally intensive.



A **DEEP NETWORK** has many layers and relatively few neurons per layer. It can achieve high levels of abstraction using relatively few neurons.

Beyond those general guidelines, however, engineers largely have to rely on experimental evidence: They run 1,000 different neural networks and simply observe which one gets the job done.

“These choices are often made by trial and error in practice,” Hanin said.

“That’s sort of a tough [way to do it] because there are infinitely many choices and one really doesn’t know what’s the best.”

A better approach would involve a little less trial and error and a little more upfront understanding of what a given neural network architecture gets you. A few papers published recently have moved the field in that direction.

“This work tries to develop, as it were, a cookbook for designing the right neural network. If you know what it is that you want to achieve out of the network, then here is the recipe for that network,” Rolnick said.

To Rope a Red Sheep

One of the earliest important theoretical guarantees about neural network architecture came three decades ago. In 1989, computer scientists proved that if a neural network has only a single computational layer, but you allow that one layer to have an unlimited number of neurons, with unlimited connections between them, the network will be capable of performing any task you might ask of it.

It was a sweeping statement that turned out to be fairly intuitive and not so useful. It’s like saying that if you can identify an unlimited number of lines in an image, you can distinguish between all objects using just one layer. That may be true in principle, but good luck implementing it in practice.

Researchers today describe such wide, flat networks as “expressive,” meaning that they’re capable in theory of capturing a richer set of connections between possible inputs (such as an image) and outputs (such as descriptions of the image). Yet these networks are extremely difficult to train, meaning it’s almost impossible to teach them how to actually produce those outputs. They’re also more computationally intensive than any computer can handle.

Boris Hanin, a mathematician at Texas A&M University, has studied the tradeoff between depth and width in neural networks.
Intel AI One Tree Studio

More recently, researchers have been trying to understand how far they can push neural networks in the other direction — by making them narrower (with fewer neurons per layer) and deeper (with more layers overall). So maybe you only need to pick out 100 different lines, but with connections for turning those 100 lines into 50 curves, which you can combine into 10 different shapes, which give you all the building blocks you need to recognize most objects.

In a [paper](#)() completed last year, Rolnick and [Max Tegmark](#)() of the Massachusetts Institute of Technology proved that by increasing depth and decreasing width, you can perform the same functions with exponentially fewer neurons. They showed that if the situation you're modeling has 100 input variables, you can get the same reliability using either 2^{100} neurons in one layer or just 2^{10} neurons spread over two layers. They found that there is power in taking small pieces and combining them at greater levels of abstraction instead of attempting to capture all levels of abstraction at once.

“The notion of depth in a neural network is linked to the idea that you can express something complicated by doing many simple things in sequence,” Rolnick said. “It’s like an assembly line.”

Rolnick and Tegmark proved the utility of depth by asking neural networks to perform a simple task: multiplying polynomial functions. (These are just equations that feature variables raised to natural-number exponents, for example $y = x^3 + 1$.) They trained the networks by showing them examples of equations and their products. Then they asked the networks to compute the products of equations they hadn’t seen before. Deeper neural networks learned the task with far fewer neurons than shallower ones.

And while multiplication isn’t a task that’s going to set the world on fire, Rolnick says the paper made an important point: “If a shallow network can’t even do multiplication then we shouldn’t trust it with anything else.”

David Rolnick, a mathematician at the University of Pennsylvania, proved that increasing a network's depth allowed a network to accomplish tasks with exponentially fewer neurons.

[Stephanie Ku\(\)](#)

Other researchers have been probing the minimum amount of width needed. At the end of September, [Jesse Johnson\(\)](#), formerly a mathematician at Oklahoma State University and now a researcher with the pharmaceutical company Sanofi, proved that at a certain point, [no amount of depth can compensate for a lack of width\(\)](#).

To get a sense of his result, imagine sheep in a field, except these are punk-rock sheep: Their wool has been dyed one of several colors. The task for your neural network is to draw a border around all sheep of the same color. In spirit, this task is similar to image classification: The network has a collection of images (which it represents as points in higher-dimensional space), and it needs to group together similar ones.

Johnson proved that a neural network will fail at this task when the width of the layers is less than or equal to the number of inputs. So for our sheep, each can be described with two inputs: an x and a y coordinate to specify its position in the field. The neural network then labels each sheep with a color and draws a border around sheep of the same color. In this case, you will need three or more neurons per layer to solve the problem.

More specifically, Johnson showed that if the width-to-variable ratio is off, the neural network won't be able to draw closed loops — the kind of loops the network would need to draw if, say, all the red sheep were clustered together in the middle of the pasture. "If none of the layers are thicker than the number of input dimensions, there are certain shapes the function will never be able to create, no matter how many layers you add," Johnson said. Papers like Johnson's are beginning to build the rudiments of a theory of neural networks. At the moment, researchers can make only very basic claims about the relationship between architecture and function — and those claims are in small proportion to the number of tasks neural networks are taking on.

So while the theory of neural networks isn't going to change the way systems are built anytime soon, the blueprints are being drafted for a new theory of how computers learn — one that's poised to take humanity on a ride with even greater repercussions than a trip to the moon.

Computers and Humans ‘See’ Differently. Does It Matter?

In some ways, machine vision is superior to human vision. In other ways, it may never catch up.

Introduction

When engineers first endeavored to teach computers to see, they took it for granted that computers would see like humans. The first proposals for computer vision in the 1960s were “clearly motivated by characteristics of human vision,” said a computer scientist at York University.

Things have changed a lot since then.

Computer vision has grown from a pie-in-the-sky idea into a sprawling field. Computers can now outperform human beings in some vision tasks, like classifying pictures — dog or wolf? — and detecting anomalies in medical images. And the way artificial “neural networks” process visual data looks increasingly dissimilar from the way humans do.

Computers are beating us at our own game by playing by different rules.

The neural networks underlying computer vision are fairly straightforward. They receive an image as input and process it through a series of steps. They first detect pixels, then edges and contours, then whole objects, before eventually producing a final guess about what they’re looking at. These are known as “feed forward” systems because of their assembly-line setup.

There is a lot we don’t know about human vision, but we know it doesn’t work like that. In our recent story, “[A Mathematical Model Unlocks the Secrets of Vision](#),” *Quanta* described a new mathematical model that tries to explain the central mystery of human vision: how the visual cortex in the brain creates vivid, accurate representations of the world based on the scant information it receives from the retina.

The model suggests that the visual cortex achieves this feat through a series of neural feedback loops that refine small changes in data from the outside world

into the diverse range of images that appear before our mind's eye. This feedback process is very different from the feed-forward methods that enable computer vision.

"This work really shows how sophisticated and in some sense different the visual cortex is" from computer vision, said [Jonathan Victor\(\)](#), a neuroscientist at Cornell University.

But computer vision is superior to human vision at some tasks. This raises the question: Does computer vision need inspiration from human vision at all?

In some ways, the answer is obviously no. The information that reaches the visual cortex is constrained by anatomy: Relatively few nerves connect the visual cortex with the outside world, which limits the amount of visual data the cortex has to work with. Computers don't have the same bandwidth concerns, so there's no reason they need to work with sparse information.

"If I had infinite computing power and infinite memory, do I need to sparsify anything? The answer is likely no," Tsotsos said.

But Tsotsos thinks it's folly to disregard human vision.

The classification tasks computers are good at today are the "low-hanging fruit" of computer vision, he said. To master these tasks, computers merely need to find correlations in massive data sets. For higher-order tasks, like scanning an object from multiple angles in order to determine what it is (think about the way you familiarize yourself with a statue by walking around it), such correlations may not be enough to go on. Computers may need to take a nod from humans to get it right.

(In an interview with [Quanta Magazine](#) last year, the artificial intelligence pioneer Judea Pearl [made this point more generally](#) when he argued that correlation training won't get AI systems very far in the long run.)

For example, a key feature of human vision is the ability to do a double take. We process visual information and reach a conclusion about what we've seen. When that conclusion is jarring, we look again, and often the second glance tells us what's really going on. Computer vision systems working in a feed-

forward manner typically lack this ability, which leads computer vision systems to [fail spectacularly at even some simple vision tasks](#).

There's another, subtler and more important aspect of human vision that computer vision lacks.

It takes years for the human visual system to mature. A 2019 paper by Tsotsos and his collaborators found that people [don't fully acquire\(\)](#) the ability to suppress clutter in a crowded scene and focus on what they're looking for until around age 17. Other research has found that the ability to perceive faces keeps improving until around age 20.

Computer vision systems work by digesting massive amounts of data. Their underlying architecture is fixed and doesn't mature over time, the way the developing brain does. If the underlying learning mechanisms are so different, will the results be, too? Tsotsos thinks computer vision systems are in for a reckoning.

"Learning in these deep learning methods is as unrelated to human learning as can be," he said. "That tells me the wall is coming. You'll reach a point where these systems can no longer move forward in terms of their development."

Computers Evolve a New Path Toward Human Intelligence

Neural networks that borrow strategies from biology are making profound leaps in their abilities. Is ignoring a goal the best way to make truly intelligent machines?

The steppingstone principle is a way to inject creativity into artificial intelligence.

Introduction

In 2007, [Kenneth Stanley\(\)](#), a computer scientist at the University of Central Florida, was playing with [Picbreeder\(\)](#), a website he and his students had created, when an alien became a race car and changed his life. On Picbreeder, users would see an array of 15 similar images, composed of geometric shapes or swirly patterns, all variations on a theme. On occasion, some might resemble a real object, like a butterfly or a face. Users were asked to select one, and they typically clicked on whatever they found most interesting. Once they did, a new set of images, all variations on their choice, would populate the screen. From this playful exploration, a catalog of fanciful designs emerged.

Stanley is a pioneer in a field of artificial intelligence called neuroevolution, which co-opts the principles of biological evolution to design smarter algorithms. With Picbreeder, each image was the output of a computational system similar to a neural network. When an image spawned, its underlying network mutated into 15 slightly different variations, each of which contributed a new image. Stanley didn't intend for Picbreeder to generate anything in particular. He merely had a hunch that he, or the public, might learn something about evolution, or about artificial intelligence.

One day Stanley spotted something resembling an alien face on the site and began evolving it, selecting a child and grandchild and so on. By chance, the round eyes moved lower and began to resemble the wheels of a car. Stanley went with it and evolved [a spiffy-looking sports car\(\)](#). He kept thinking about the fact that if he had started trying to evolve a car from scratch, instead of from an alien, he might never have done it, and he wondered what that implied about attacking problems directly. "It had a huge impact on my whole life," he said. He looked at other interesting images that had emerged on Picbreeder, traced their lineages, and realized that nearly all of them had evolved by way of something that looked completely different. "Once I saw the evidence for that, I was just blown away."

Stanley's realization led to what he calls the steppingstone principle — and, with it, a way of designing algorithms that more fully embraces the endlessly creative potential of biological evolution.

Evolutionary algorithms have been around for a long time. Traditionally, they've been used to solve specific problems. In each generation, the solutions that perform best on some metric — the ability to control a two-legged robot, say — are selected and produce offspring. While these algorithms have seen some successes, they can be more computationally intensive than other approaches such as “[deep learning](#),” which has exploded in popularity in recent years.

The steppingstone principle goes beyond traditional evolutionary approaches. Instead of optimizing for a specific goal, it embraces creative exploration of all possible solutions. By doing so, it has paid off with groundbreaking results. Earlier this year, one system based on the steppingstone principle mastered two video games that had stumped popular machine learning methods. And in [a paper published last week in Nature\(\)](#), DeepMind — the artificial intelligence company that pioneered the use of deep learning for problems such as the game of Go — reported success in combining deep learning with the evolution of a diverse population of solutions.

The steppingstone's potential can be seen by analogy with biological evolution. In nature, the tree of life has no overarching goal, and features used for one function might find themselves enlisted for something completely different. Feathers, for example, likely [evolved for insulation\(\)](#) and only later became handy for flight.

Biological evolution is also the only system to produce human intelligence, which is the ultimate dream of many AI researchers. Because of biology's track record, Stanley and others have come to believe that if we want algorithms that can navigate the physical and social world as easily as we can — or better! — we need to imitate nature's tactics. Instead of hard-coding the rules of reasoning, or having computers learn to score highly on specific performance metrics, they argue, we must let a population of solutions blossom. Make them prioritize novelty or interestingness instead of the ability to walk or talk. They may discover an indirect path, a set of steppingstones, and wind up walking and talking better than if they'd sought those skills directly.

New, Interesting, Diverse

After Picbreeder, Stanley set out to demonstrate that neuroevolution could overcome the most obvious argument against it: “If I run an algorithm that’s creative to such an extent that I’m not sure what it will produce,” he said, “it’s very interesting from a research perspective, but it’s a harder sell commercially.”

He hoped to show that by simply following ideas in interesting directions, algorithms could not only produce a diversity of results, but solve problems. More audaciously, he aimed to show that completely ignoring an objective can get you there faster than pursuing it. He did this through an approach called [novelty search\(\)](#).

The system started with a neural network, which is an arrangement of small computing elements called neurons connected in layers. The output of one layer of neurons gets passed to the next layer via connections that have various “weights.” In a simple example, input data such as an image might be fed into the neural network. As the information from the image gets passed from layer to layer, the network extracts increasingly abstract information about its contents. Eventually, a final layer calculates the highest-level information: a label for the image.

For Kenneth Stanley, a computer scientist at Uber AI Labs and the University of Central Florida, the steppingstone principle explains innovation.

Asa Mathat

In neuroevolution, you start by assigning random values to the weights between layers. This randomness means the network won’t be very good at its job. But from this sorry state, you then create a set of random mutations — offspring neural networks with slightly different weights — and evaluate their abilities. You keep the best ones, produce more offspring, and repeat. (More advanced neuroevolution strategies will also introduce mutations in the number and arrangement of neurons and connections.) Neuroevolution is a meta-algorithm, an algorithm for designing algorithms. And eventually, the algorithms get pretty good at their job.

To test the steppingstone principle, Stanley and his student [Joel Lehman\(\)](#) tweaked the selection process. Instead of selecting the networks that performed best on a task, novelty search selected them for how different they were from the ones with behaviors most similar to theirs. (In Picbreeder,

people rewarded interestingness. Here, as a proxy for interestingness, novelty search rewarded novelty.)

In one test, they placed virtual wheeled robots in a maze and evolved the algorithms controlling them, hoping one would find a path to the exit. They ran the evolution from scratch 40 times. A comparison program, in which robots were selected for how close (as the crow flies) they came to the exit, evolved a winning robot only 3 out of 40 times. Novelty search, which completely ignored how close each bot was to the exit, succeeded 39 times. It worked because the bots managed to avoid dead ends. Rather than facing the exit and beating their heads against the wall, they explored unfamiliar territory, found workarounds, and won by accident. “Novelty search is important because it turned everything on its head,” said [Julian Togelius\(\)](#), a computer scientist at New York University, “and basically asked what happens when we don’t have an objective.”

Once Stanley had made his point that the pursuit of objectives can be a hindrance to reaching those objectives, he looked for clever ways to combine novelty search and specific goals. That led him and Lehman to create a system that mirrors nature’s evolutionary niches. In [this approach\(\)](#), algorithms compete only against others that are similar to them. Just as worms don’t compete with whales, the system maintains separate algorithmic niches from which a variety of promising approaches can emerge.

[Such evolutionary algorithms with localized competition\(\)](#) have shown proficiency at [processing pixels, controlling a robot arm\(\)](#), and ([as depicted on the cover of Nature\(\)](#)) helping a six-legged robot quickly adapt its gait after losing a limb, the way an animal would. A key element of these algorithms is that they foster steppingstones. Instead of constantly prioritizing one overall best solution, they maintain a diverse set of vibrant niches, any one of which could contribute a winner. And the best solution might descend from a lineage that has hopped between niches.

Evolved to Win

For Stanley, who is now at Uber AI Labs, the steppingstone principle explains innovation: If you went back in time with a modern computer and told people developing vacuum tubes to abandon them and focus on laptops, we’d have neither. It also explains evolution: We evolved from flatworms, which were not especially intelligent but did have bilateral symmetry. “It’s totally unclear that the discovery of bilateral symmetry had anything to do with intelligence, let alone with Shakespeare,” Stanley said, “but it does.”

Neuroevolution itself has followed an unexpectedly circuitous path over the past decade. For a long time, it has lived in the shadows of other forms of AI.

One of its biggest drawbacks, according to [Risto Miikkulainen\(\)](#), a computer scientist at the University of Texas, Austin (and Stanley's former doctoral adviser), is the amount of computation it requires. In traditional machine learning, as you train a neural network, it gradually gets better and better. With neuroevolution, the weights change randomly, so the network's performance might degrade before it improves.

Another drawback is the basic fact that most people have a particular problem that they'd like to solve. A search strategy that optimizes for interestingness might get you to a creative solution for that particular problem. But it could lead you astray before it puts you on the right path.

Then again, no strategy is perfect. In the past five years or so, research has exploded in different areas of AI research such as deep learning and reinforcement learning. In reinforcement learning, an algorithm interacts with the environment — a robot navigates the real world, or a player competes in a game — and learns through trial and error which behaviors lead to desired outcomes. Deep reinforcement learning was used by DeepMind to create a program that could [beat the world's best players at Go](#), a feat that many thought was still years or decades away.

But reinforcement learning can get stuck in a rut. Sparse or infrequent rewards don't give algorithms enough feedback to enable them to proceed toward their goal. Deceptive rewards — awarded for short-term gains that hinder long-term progress — trap algorithms in dead ends. So while reinforcement learning can whip humans at Space Invaders or Pong — games with frequent points and clear goals — they've fallen flat in other classic games that lack those features.

Within the past year, AI based on the steppingstone principle finally managed to crack a number of long-standing challenges in the field.

In the game Montezuma's Revenge, Panama Joe navigates from room to room in an underground labyrinth, collecting keys to open doors while avoiding enemies and obstacles like snakes and fire pits. To beat the game, Stanley, Lehman, [Jeff Clune\(opens a new tab\)](#), [Joost Huizinga\(\)](#) and [Adrien Ecoffet\(\)](#), all

five working at Uber AI Labs, developed a system where Panama Joe essentially wanders around and randomly attempts various actions. Each time he reaches a new game state — a new location with a new set of possessions — he files it away in his memory, along with the set of actions he took to get there. If he later finds a faster path to that state, it replaces the old memory. During training, Panama Joe repeatedly picks one of those stored states, explores randomly for a bit, and adds to his memory any new states he finds.

Eventually one of those states is the state of winning the game. And Panama Joe has in his memory all the actions he took to get there. He's done it with no neural network or reinforcement learning — no rewards for collecting keys or nearing the labyrinth's end — just random exploration and a clever way to collect and connect steppingstones. This approach managed to beat not only the best algorithms but also the human world record for the game.

The same technique, which the researchers call [Go-Explore\(\)](#), was used to beat human experts at Pitfall!, a game where Pitfall Harry navigates a jungle in search of treasure while avoiding crocodiles and quicksand. No other machine learning AI had scored above zero.

Now even DeepMind, that powerhouse of reinforcement learning, has revealed its growing interest in neuroevolution. In January, the team showed off [AlphaStar\(\)](#), software that can beat top professionals at the complex video game StarCraft II, in which two opponents control armies and build colonies to dominate a digital landscape. AlphaStar evolved a population of players that competed against and learned from each other. In last week's [Nature paper\(\)](#), DeepMind researchers announced that an updated version of AlphaStar has been ranked among the top 0.2% of active StarCraft II players on a popular gaming platform, becoming the first AI to reach the top tier of a popular esports without any restrictions.

Jeff Clune, a computer scientist at Uber AI Labs and the University of Wyoming, argues that open-ended discovery is likely the fastest way to achieve humanlike artificial intelligence.

Wayne Cunningham, Uber

“For a long time with the AlphaStar agents, they were getting better, but they were always exploitable,” said [Max Jaderberg\(\)](#), a computer scientist at DeepMind who worked on the project. “You would train an agent, and it would have very, very good performance on average, but you could always train something against this agent and find holes in this agent.”

As in the children's game rock-paper-scissors, there is no single best game strategy in StarCraft II. So DeepMind encouraged its population of agents

to [evolve a diversity of strategies](#)() — not as steppingstones but as an end in itself. When AlphaStar beat two pros each five games to none, it combined the strategies from five different agents in its population. The five agents had been chosen so that not all of them would be vulnerable to any one opponent strategy. Their strength was in their diversity.

AlphaStar demonstrates one of the main uses of evolutionary algorithms: maintaining a population of different solutions. Another recent DeepMind project shows the other use: optimizing a single solution. Working with Waymo, Alphabet’s autonomous car project, the team [evolved algorithms for identifying pedestrians](#)(). To avoid getting stuck with an approach that works fairly well, but that isn’t the best possible strategy, they maintained “niches” or subpopulations, so that novel solutions would have time to develop before being crushed by the established top performers.

Population-based algorithms have become more popular in recent years, in part because “they’re a good match to the type of compute that we have now,” said [Raia Hadsell](#)(), a research scientist and head of robotics at DeepMind, using an industry standard term for computing resources. Hadsell invited Clune, Lehman and Stanley to give [a two-hour presentation](#)() of their work at the International Conference on Machine Learning in June. “I think it’s an important area of research for AI,” she said, “because it is complementary to the deep learning approaches that have driven the field.”

AI That Designs AI

All of the algorithms discussed so far are limited in their creativity. AlphaStar can only ever come up with new StarCraft II strategies. Novelty search can find novelty within only one domain at a time — solving a maze or walking a robot.

Biological evolution, on the other hand, produces endless novelty. We have bacteria and kelp and birds and people. That’s because solutions evolve, but so do problems. The giraffe is a response to the problem of the tree. Human innovation proceeds likewise. We create problems for ourselves — could we put a person on the moon? — and then solve them.

Decades of research have taught us that these algorithms constantly surprise us and outwit us.

[Jeff Clune](#)

To mirror this open-ended conversation between problems and solutions, earlier this year Stanley, Clune, Lehman and another Uber colleague, Rui Wang, released an algorithm called [POET\(\)](#), for Paired Open-Ended Trailblazer. To test the algorithm, they evolved a population of virtual two-legged bots. They also evolved a population of obstacle courses for the bots, complete with hills, trenches and tree stumps. The bots sometimes traded places with each other, attempting new terrain. For example, one bot learned to cross flat terrain while dragging its knee. It was then randomly switched to a landscape with short stumps, where it had to learn to walk upright. When it was switched back to its first obstacle course, it completed it much faster. An indirect path allowed it to improve by taking skills learned from one puzzle and applying them to another.

POET could potentially design new forms of art or make scientific discoveries by inventing new challenges for itself and then solving them. It could even go much further, depending on its world-building ability. Stanley said he hopes to build algorithms that could still be doing something interesting after a billion years.

Evolution “invented sight, it invented photosynthesis, it invented human-level intelligence, it invented all of it, and all of it in one run of an algorithm,” Stanley said. “To capture one tiny iota of that process I think could be incredibly powerful.”

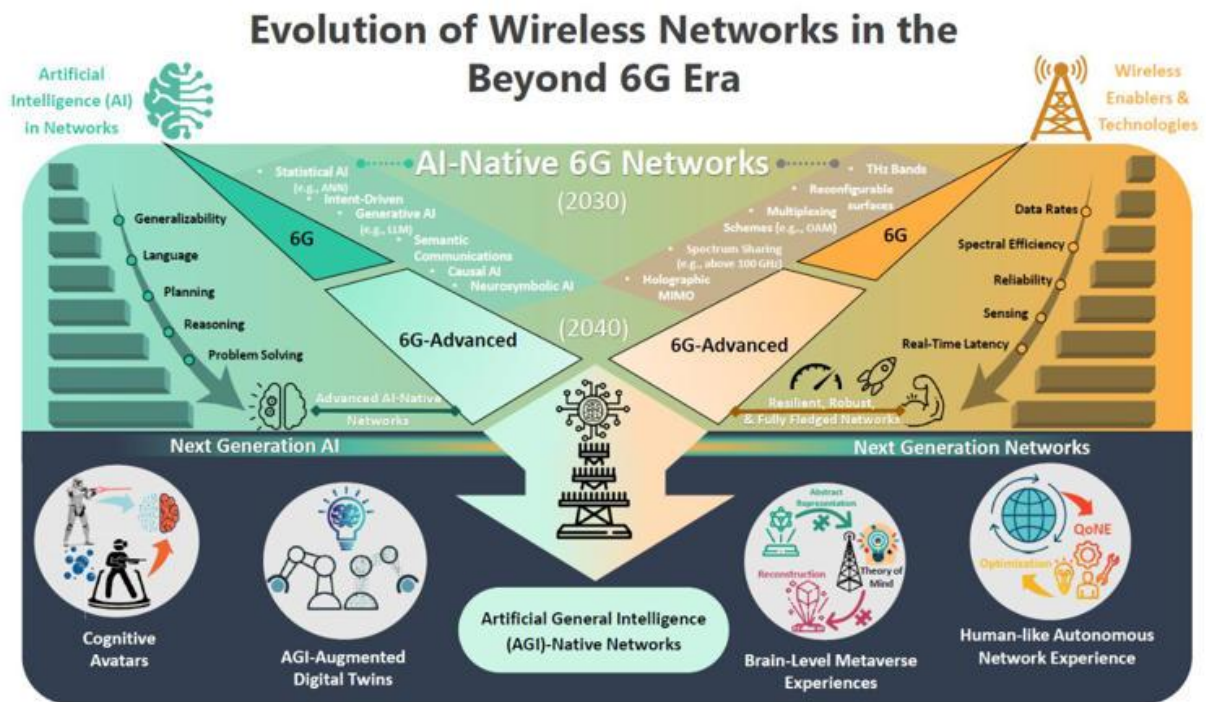
In [a recent paper\(\)](#), Clune argues that [open-ended discovery\(\)](#) is likely the fastest path toward artificial general intelligence — machines with nearly all the capabilities of humans. Most of the AI field is focused on manually designing all the building blocks of an intelligent machine, such as different types of neural network architectures and learning processes. But it’s unclear how these might eventually get bundled together into a general intelligence.

Instead, Clune thinks more attention should be paid to AI that designs AI. Algorithms will design or evolve both the neural networks and the environments in which they learn, using an approach like POET’s. Such open-ended exploration might lead to human-level intelligence via paths we never would have anticipated — or a variety of alien intelligences that could teach us a lot about intelligence in general. “Decades of research have taught us that these algorithms constantly surprise us and outwit us,” he said. “So it’s completely hubristic to think that we will know the outcome of these processes, especially as they become more powerful and open-ended.”

It may also be hubristic to exert too much control over researchers. A sweet irony to Stanley's story is that he originally pitched Picbreeder to the National Science Foundation, which rejected his grant application, saying its objective wasn't clear. But the project has led to papers, talks, a [book\(\)](#) and a startup — Geometric Intelligence, [purchased\(\)](#) by Uber to form the core of Uber AI Labs. "To me," Stanley said, "one of the things that's really kind of striking and maybe crazy is that the story of how I got here is basically the same as the algorithmic insight that got me here. The thing that led me to the insight is actually explained by the insight itself."

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

Researchers publish blueprint to fuse wireless technologies and AI



Credit: Illustration courtesy of Walid Saad, Omar Hashash, and Christo Thomas.

There's a major difference between humans and current artificial intelligence (AI) capabilities: common sense. According to a new paper by Walid Saad, professor in the College of Engineering and the Next-G Wireless Lead at the Virginia Tech Innovation Campus, a true revolution in wireless technologies is only possible through endowing the system with the next generation of AI that can think, imagine, and plan akin to humans.

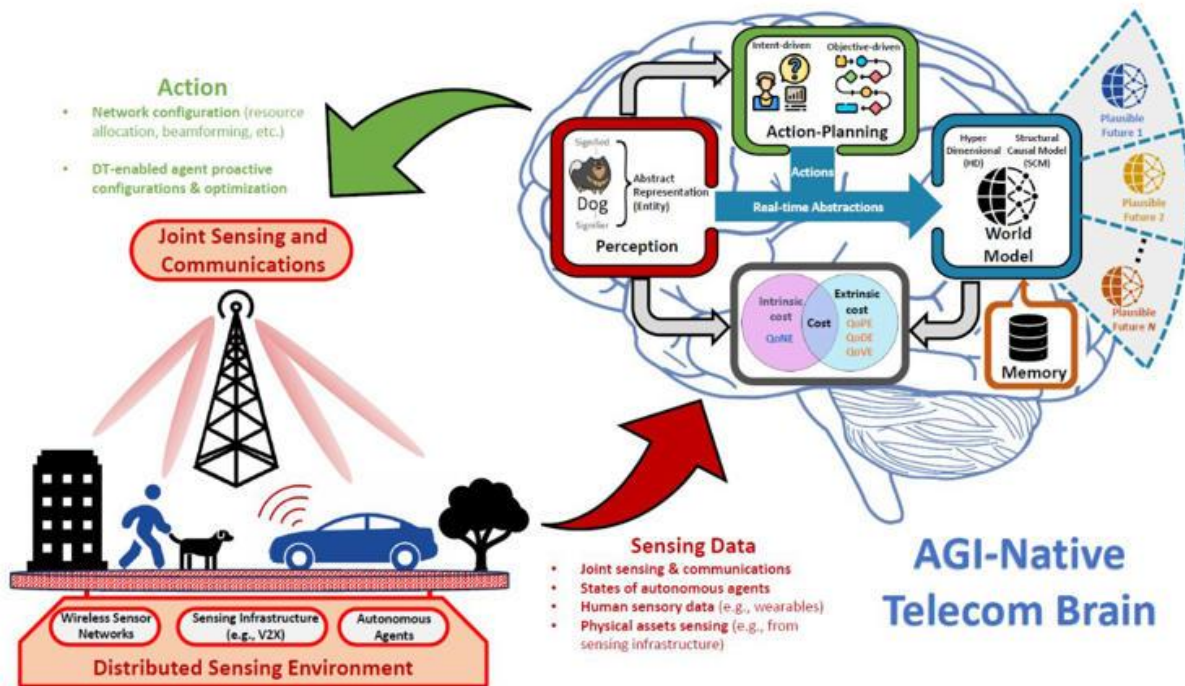
Published in the *Proceedings of the IEEE* journal's Special Issue on the Road to 6G with Ph.D. student Omar Hashash and postdoctoral associate Christo Thomas, the paper's findings suggest:

- The missing link in the wireless revolution is next-generation AI.
- The missing link in the next generation of AI is wireless technologies.
- The solution is to bring AI closer to human intelligence through common sense.

"We're looking at least 10 or 15 years down the line before we have a wireless network with artificial general intelligence [AGI] that can think, plan, and imagine," said Saad, who is a professor in the Bradley Department of Electrical and Computer Engineering.

"We have a blueprint and concrete road map. The entire vision might not be immediately deployable, but pieces of it can be implemented now. We're trying to position this paper in a way to tell the community that there is a path to something really revolutionary—step by step we can work toward a living, thinking wireless network."

Previous generations of wireless networks have been defined by several enhancements to core components, such as new antennas and communication technologies that have improved performance. According to the researchers, not even the leap from 5G to 6G, characterized by the addition of an AI-architecture embedded in wireless systems and an open radio access network, will be revolutionary enough to meet future processing and networking needs.



Credit: Illustration courtesy of Walid Saad, Omar Hashash, and Christo Thomas.

"That is where things start to become thrilling," Hashash said. "The next generation of wireless networks and AI are converging hand in hand, but few are seeing how they can actually be seamlessly merged."

Physical networks endowed with AI

At first, Saad, Hashash, and Thomas were [focusing on the metaverse](#) and building on what is currently being explored in 6G: embedding AI across wireless systems, referred to as AI-native networks.

"The problem is researchers are using classical AI tools that are designed for other tasks such as computer vision, which makes them limited in many ways when it comes to communication networks," Saad said. "To fuse the real world with a virtual world, you basically have to mirror it. This is not something that old school AI can do."

Although humans develop common sense through a world model and understanding the intuitive physics of the environment, current AI systems are trained on data. They tend to extract patterns and capture underlying correlative

relationships, but these AI systems cannot reasonably navigate unforeseen scenarios.

In its next phase, 6G hopes to overcome this narrow, statistical, rule-based AI solution so that the network improves its sustainability, generalizability, trustworthiness, and explainability. To date, we still do not have an AI system that can deal with new and unfamiliar scenarios, because they lack an important human trait: common sense.

"Common sense allows us to deal with new scenarios, learn by analogy, and connect the dots to fill in missing plausible elements when needed," Saad said. "Simply put, the current level of AI is good at extracting statistical relationships from data, but it's very bad at reasoning and generalizing to novel, unexpected situations—things that most humans master perfectly."

To blend the physical, virtual, and digital dimensions seamlessly—for example, to don a virtual reality headset and "travel" across space and time from the comfort of home—the next generation of wireless systems will need extreme wireless quality-of-service requirements for perfectly synchronizing worlds. It also will need a hyper form of AI that can enable the network to seamlessly orchestrate the physical, virtual, and digital dimensions, something that only a real human-like network can achieve.

In a nutshell, one of the challenges for the next generation of wireless beyond 6G is not only the physical constraints of the wireless technologies, but the limited capacities of current AI technologies.

The telecom brain

As the research continued, team members were surprised to find they were not only building on their previous wireless studies, but that their research converged with optimistic advancements in AI toward human-level intelligence.

"On the one hand, the metaverse with its digital world can enable a real-time perception of the physical world, which is an essential factor to enable AGI-native networks," Hashash said. "On the other hand, the metaverse promises to bring forth novel use cases and applications such as cognitive avatars that require common sense abilities."

The metaverse, with emerging applications such as digital twins and its ability to have identical digital representations of the physical world, could provide crucial opportunities for networks to acquire perception, hyperdimensional world models, planning abilities, and analogical reasoning. This architecture would provide the missing link that would make it a real "brain," equipping the network to handle unforeseen obstacles and predict new scenarios outside of its training data.

"We must create wireless networks with intrinsic abilities to understand the mathematical mechanics behind the designed AI models, the physical properties of real-world objects, and their interactions with each other," Thomas said.

"This requires us to fuse mathematical principles, category theory, and neuroscience to model the physical world and understand the complex operations of the human brain. We are therefore advocating for revisiting the fundamentals in AI, wireless networks, mathematics, and neuroscience."

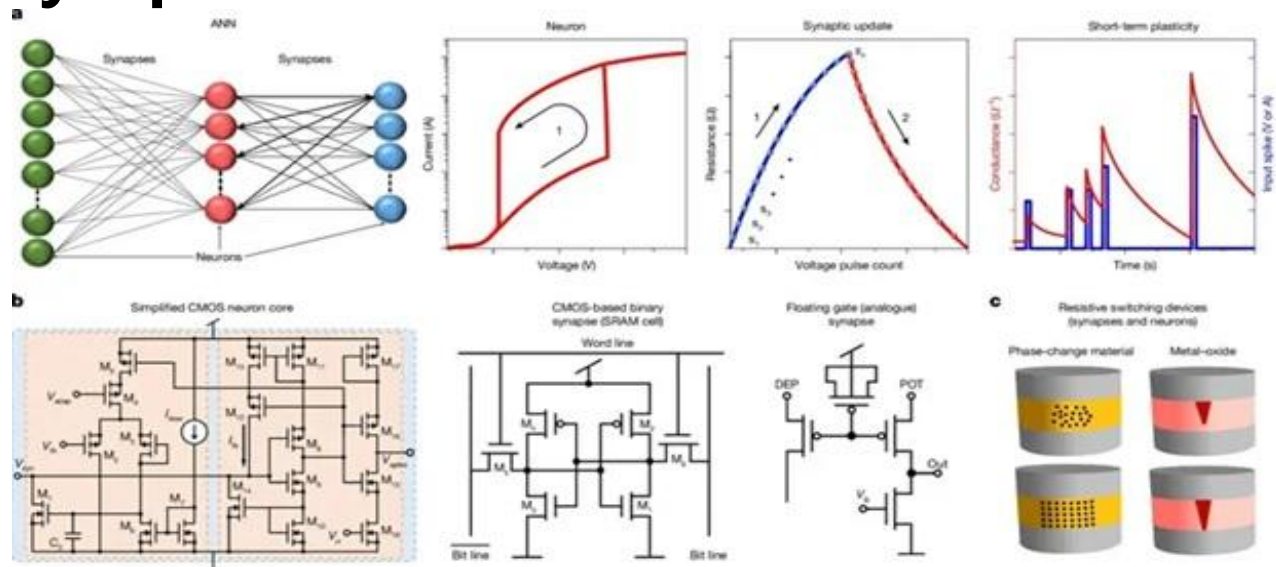
Instead of incremental advances to known wireless technologies, the researchers propose a paradigm shift. This shift would surpass the AI-native wireless system anticipated with 6G and aspire to a system that is equipped with human levels of intelligence—intelligence that is learned at the intersection of the digital world and future wireless networks. Then, this AGI-native network is set to endow some of its common sense abilities to digital twins, thereby unleashing a new breed of human-like AI agents.

"The missing link is really the wireless network and its components like digital twins, because we can use a twin that exists as a basis for a world model thereby enabling human-level-like thinking and integrating these 'thought' processes in the wireless network now," Saad said. "We could actually overcome some of the current network limitations, unleashing a completely new era of wireless networking. It is a win-win strategy for the wireless and AI evolutions."

More information: Walid Saad et al, Artificial General Intelligence (AGI)-Native Wireless Systems: A Journey Beyond 6G, *Proceedings of the IEEE* (2025). DOI: [10.1109/JPROC.2025.3526887](https://doi.org/10.1109/JPROC.2025.3526887)

Provided by Virginia Tech

Advancing semiconductor devices for AI: Single transistor acts like neuron and synapse



Transistors and their use in neuro-synaptic-mimicking devices. Credit: Nature (2025). DOI: 10.1038/s41586-025-08742-4

Researchers from the National University of Singapore (NUS) have demonstrated that a single, standard silicon transistor, the fundamental building block of microchips used in computers, smartphones and almost every electronic system, can function like a biological neuron and synapse when operated in a specific, unconventional way.

Led by Associate Professor Mario Lanza from the Department of Materials Science and Engineering at the College of Design and Engineering, NUS, the research team's work presents a highly scalable and energy-efficient solution for hardware-based artificial neural networks (ANNs).

This brings neuromorphic computing—where chips could process information more efficiently, much like the human brain—closer to reality. Their study was [published](#) in the journal *Nature*.

Putting the brains in silicon

The world's most sophisticated computers already exist inside our heads. Studies show that the human brain is, by and large, more energy-efficient than electronic processors, thanks to almost 90 billion neurons that form some 100 trillion connections with each other, and synapses that tune their strength over time—a process known as synaptic plasticity, which underpins learning and memory.

For decades, scientists have sought to replicate this efficiency using artificial neural networks (ANNs). ANNs have recently driven remarkable advances in artificial intelligence (AI), loosely inspired by how the brain processes information.

But while they borrow biological terminology, the similarities run only skin deep—software-based ANNs, such as those powering large language models like ChatGPT, have a voracious appetite for computational resources, and hence, electricity. This makes them impractical for many applications.

Neuromorphic computing aims to mimic the computing power and energy efficiency of the brain. This requires not only redesigning system architecture to carry out memory and computation at the same place—the so-called in-memory computing (IMC)—but also to develop electronic devices that exploit physical and electronic phenomena capable of replicating more faithfully how neurons and synapses work.

However, current neuromorphic computing systems are stymied by the need for complicated multi-transistor circuits or emerging materials that are yet to be validated for large-scale manufacturing.

"To enable true neuromorphic computing, where microchips behave like biological neurons and synapses, we need hardware that is both scalable and energy-efficient," said Professor Lanza.

The NUS research team has now demonstrated that a single, standard silicon transistor, when arranged and operated in a specific way, can replicate both neural firing and synaptic weight changes—the fundamental mechanisms of biological neurons and synapses.

This was achieved through adjusting the resistance of the bulk terminal to specific values, which allows controlling two physical phenomena taking place in

the transistor: punch through impact ionization and charge trapping. Moreover, the team built a two-transistor cell capable of operating either in a neuron or a synaptic regime, which the researchers have called "Neuro-Synaptic Random Access Memory," or NS-RAM.

"Other approaches require complex transistor arrays or novel materials with uncertain manufacturability, but our method makes use of commercial CMOS (complementary metal-oxide-semiconductor) technology, the same platform found in modern computer processors and memory microchips," explained Professor Lanza.

"This means it's scalable, reliable and compatible with existing semiconductor fabrication processes."

Through experiments, the NS-RAM cell demonstrated low power consumption, maintained stable performance over many cycles of operation and exhibited consistent, predictable behavior across different devices—all of which are desired attributes for building reliable ANN hardware suited for real-world applications.

The team's breakthrough marks a step change in the development of compact, power-efficient AI processors that could enable faster, more responsive computing.

More information: Sebastian Pazos et al, Synaptic and neural behaviours in a standard silicon transistor, *Nature* (2025). DOI: [10.1038/s41586-025-08742-4](https://doi.org/10.1038/s41586-025-08742-4)

Provided by National University of Singapore

An AI Model Has Officially Passed the Turing Test



OpenAI's GPT-4.5 model passed a Turing Test with flying colors, and even came off as human more than the actual humans.

One of the industry's leading large language models has passed a Turing test, a longstanding barometer for human-like intelligence.

In a new [preprint study](#) awaiting peer review, researchers report that in a three-party version of a Turing test, in which participants chat with a human and an AI at the same time and then evaluate which is which, OpenAI's GPT-4.5 model was deemed to be the human 73 percent of the time when it was instructed to adopt a persona. That's significantly higher than a random chance of 50 percent, suggesting that the Turing test has resoundingly been beaten.

The research also evaluated Meta's LLaMa 3.1-405B model, OpenAI's [GPT-4o model](#), and an early chatbot known as [ELIZA](#) developed some eighty years ago.

"People were no better than chance at distinguishing humans from GPT-4.5 and LLaMa (with the persona prompt)," wrote lead author Cameron Jones, a researcher at UC San Diego's Language and Cognition Lab, in an [X thread about the work](#). "And 4.5 was even judged to be human significantly *more* often than actual humans!"

The Turing test is named after British mathematician and computer scientist Alan Turing. In 1950, Turing proposed that one way to assess a machine's intelligence was by having it engage in text-based conversations with a human interrogator, who at the same time would hold a text-based conversation with another human, out of sight. Turing called this the "imitation game." If the interrogator couldn't correctly determine which respondent was the computer and which was the human, it would suggest, on a very general level, that the machine could think like a human.

In this latest study, the researchers carried out the famed experiment on an online hub. For eight rounds, a pool of nearly 300 participants were randomly assigned to either be an interrogator or one of the two "witnesses" being interrogated, with the other "witness" being a chatbot.

A key point here is how the AI models were prompted. One type was a "no-persona" prompt, in which the AI was given only basic instructions: "You are about to participate in a Turing test. Your goal is to convince the interrogator that you are a human."

For the "persona" prompt, on the other hand, the AI was specifically told to put on a specific persona, like a young person who's knowledgeable about the internet and culture.

These instructions made a world of difference. Without persona prompting, GPT-4.5 achieved an overall win rate of merely 36 percent, significantly down from its Turing-trumping 73 percent. As a baseline, GPT-4o, which powers the current version of ChatGPT and only received no-persona prompts, achieved an even less convincing 21 percent. (Somehow, the [ancient ELIZA](#) marginally surpassed OpenAI's flagship model with a 23 percent success rate.)

The results are intriguing. But as vaunted as the Turing test has become in AI and philosophy circles, it's not unequivocal proof that an AI thinks like we do.

"It was not meant as a literal test that you would actually run on the machine — it was more like a thought experiment," François Chollet, a software engineer at Google, [told Nature](#) in 2023.

For all their faults, LLMs are master conversationalists, trained on unfathomably vast sums of human-composed text. Even faced with a question they don't

understand, an LLM will weave a plausible-sounding response. It's becoming clearer and clearer that AI chatbots are excellent at mimicking us — so perhaps assessing their wits with an "imitation game" is becoming a bit of a moot point.

As such, Jones doesn't think the implications of his research — whether LLMs are intelligent like humans — are clear-cut.

"I think that's a very complicated question..." Jones [tweeted](#). "But broadly I think this should be evaluated as one among many other pieces of evidence for the kind of intelligence LLMs display."

"More pressingly, I think the results provide more evidence that LLMs could substitute for people in short interactions without anyone being able to tell," he added. "This could potentially lead to automation of jobs, improved social engineering attacks, and more general societal disruption."

Jones closes out by emphasizing that the Turing test doesn't just put the machines under the microscope — it also reflects humans' ever-evolving perceptions of technology. So the results aren't static: perhaps as the public becomes more familiar with interacting with AIs, they'll get better at sniffing them out, too.

"Thinking" Visually Boosts AI Problem Solving

Giving language models a “mind’s eye” helps them tackle spatial reasoning tasks

EDD GENT

12 FEB 2025

4 MIN READ

Edd Gent is a Contributing Editor for IEEE Spectrum.

The model's "mind's eye" visualization of the scene (right) matched the reality of a maze-like game.

CHENGZU LI, WENSHAN WU ET AL./GYMNASIUM

ARTIFICIAL INTELLIGENCE | NEURAL NETWORKS | MULTIMODAL ROBOTS

When humans try to solve problems, they often visualize the tasks in their heads. New research suggests that enabling [artificial intelligence](#) to do the same could boost performance on spatial reasoning challenges.

While [large language models](#) excel at many text-based tasks, they often struggle with those that require more complex reasoning. One of the most promising approaches for boosting their performance on these kinds of problems is a technique known as “chain-of-

thought” (CoT) [prompting](#), where users ask the model to “think” through them step-by-step.

This can lead to significant improvements on various reasoning tasks, especially in [mathematics](#), coding, and logic. But the language-focused technique has proved less effective for problems requiring spatial or visual reasoning. To try and close that gap, researchers at the [University of Cambridge](#) and [Microsoft Research](#) have developed a new approach that lets AI “think” in both text and images.

The technique enables multimodal [large language models](#), which can process both image and text data, to generate visual representations of their intermediate reasoning steps. In non-peer reviewed [research posted to arXiv](#), the researchers report that when they tested the approach on spatial reasoning challenges involving 2D mazes, they saw significant improvements over the typical CoT technique on the most challenging scenarios.

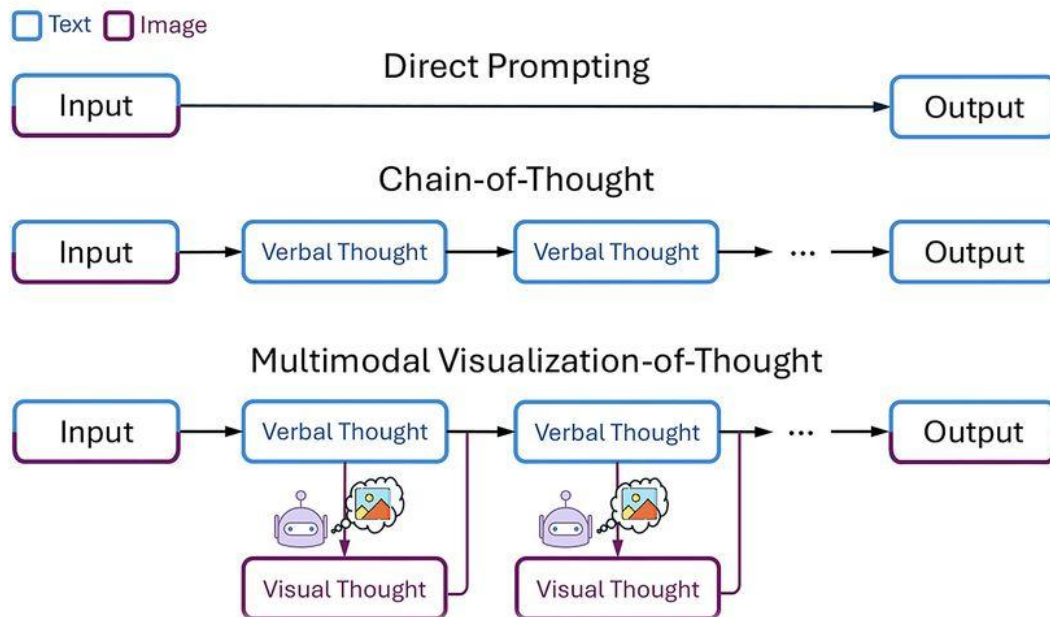
“Spatial relations and layouts and also some geometric features are very hard to describe with pure text,” says co-lead author [Chengzu Li](#), a Ph.D. student at Cambridge. “That’s why we think that reasoning with pure text would limit the performance of the model in spatial tasks. And that’s the main motivation for introducing visual ‘thoughts,’” he says.

How AI Visual Reasoning Works

This is not the first attempt to allow AI to reason visually. But Li says previous approaches have either involved extracting information from images and converting it to text before reasoning with it, or have relied on external software tools or specialized vision models to enable visual reasoning.

The new approach enables a single multimodal model to generate both visual and text reasoning steps itself. This work only recently became feasible, says Li, thanks to the development of more powerful multimodal AI. Older models could interpret images and text, but could only generate text outputs. For these experiments, the researchers used a model called [Anole](#) that can respond in either modality.

This model is an open-source extension of Meta's [Chameleon](#) multimodal model: the researchers behind Anole retrained it to generate sequences of text interleaved with images. For instance, it can generate a step-by-step recipe with an image for each step. Li and colleagues took this pre-trained model and fine-tuned it on text and image data from three maze-like games with different levels of complexity. They called their fine-tuned version Multimodal Visualization of Thought (MVoT).



The researchers tested the new technique (bottom), which generates both visual and verbal thoughts, against one that reasons only in text (middle) and one that skips reasoning and jumps straight to the answer.[CHENGZU LI, WENSHAN WU ET AL.](#)

The goal for the model was to work out what would happen if it took a pre-determined series of actions in each maze. During training, the model was shown examples that included images of the starting position in the maze and a textual description of the task, a series of reasoning steps featuring text descriptions of actions and images of where the player is on the map, and finally an answer as to what the outcome would be for those actions, such as reaching the desired destination or falling down a hole. During testing the model was only given the starting image and a sequence of actions to perform. It then generated

image and text reasoning steps followed by a prediction of what would happen.

The researchers compared MVoT to four other models, three of which they fine-tuned themselves. The first two versions of the model were trained only on text data regarding the maze: One model jumped straight from a prompt to generating a final answer, the other used textual CoT reasoning. Another model was trained on examples of both image and text reasoning, but then did its own reasoning purely in text. Finally, they compared MVoT's performance on the maze tasks to that of the GPT-4o model from [OpenAI](#), which is the company's most advanced multimodal model.

They found that on all three games, the MVoT model significantly outperformed all models apart from the one using traditional text CoT. That model actually did slightly better on the two simpler mazes, successfully predicting the outcome 98 percent of the time on both, compared to MVoT's scores of 93 percent and 95 percent. But the traditional text CoT model did much worse on the most complicated game, scoring just 61 percent compared to MVoT's 86 percent. They tested both models on progressively larger mazes and while MVoT's

performance remained stable, the other model's performance plummeted as maze size increased.

The researchers say this outcome is likely because CoT relies on accurate textual descriptions of the environment, which get harder the more complex the mazes become. In contrast, the inclusion of images in the reasoning process appears to make MVoT much better at dealing with more challenging environments.

Applications for AI Visual Reasoning

While the tests the researchers used are simple, Li says extending this approach into more complex domains could have broad applications. One of the most compelling is [robotics](#), where the approach could help machines reason more effectively about the visual input they get from the environment. It could also help AI tutors better illustrate and explain ideas, particularly in areas like geometry. More broadly, he says the approach can boost model interpretability by giving humans a clear picture of what the model is thinking about in spatial tasks.

One potential gap, admits Li, is that the model has no mechanism for deciding when to reason visually or when to

reason via text. At present, the model simply alternates between the two, which works well for these maze navigation challenges that have discrete steps but may be less appropriate for more complex spatial reasoning tasks.

“We haven’t really touched on when is the appropriate time to do a visual reasoning process or not,” Li says. “But I think it’s definitely one of the very interesting directions to further explore.” One possibility, he adds, would be to generate reasoning sequences with both visual and text descriptions at each step, and then get humans to provide feedback on which is more expressive. This feedback could then be used to train the model to pick the best option at each reasoning step.

How neural networks represent data: A potential unifying theory for key deep learning phenomena



How do neural networks work? It's a question that can confuse novices and experts alike. A team from MIT's Computer Science and Artificial Intelligence Lab (CSAIL) says that understanding these representations, as well as how they inform the ways that neural networks learn from data, is crucial for improving the interpretability, efficiency, and generalizability of deep learning models.

With that mind, the CSAIL researchers have developed a new framework for understanding how representations form in neural networks. Their Canonical Representation Hypothesis (CRH) posits that, during training, neural networks inherently align their latent representations, weights, and neuron gradients within each layer. This alignment implies that neural networks naturally learn compact representations based on the degree and modes of deviation from the CRH.

Senior author Tomaso Poggio says that, by understanding and leveraging this alignment, engineers can potentially design networks that are more efficient and easier to understand. The research is [posted](#) to the *arXiv* preprint server.

The team's corresponding Polynomial Alignment Hypothesis (PAH) posits that, when the CRH is broken, distinct phases emerge in which the representations, gradients, and weights become polynomial functions of each other. Poggio says that the CRH and PAH offer a potential unifying theory for key deep learning phenomena such as neural collapse and the neural feature ansatz (NFA).

The new CSAIL paper about the project provides experimental results across various settings to support the CRH and PAH on tasks that include image classification and self-supervised learning. The CRH suggests possibilities for manually injecting noise into neuron gradients to engineer specific structures in the model's representations. Poggio says that a key future direction is to understand the conditions that lead to each phase and how these phases affect the behavior and performance of models.

"The paper offers a new perspective on understanding the formation of representations in neural networks through the CRH and PAH," says Poggio. "This provides a framework for unifying existing observations and guiding future research in deep learning."

Co-author Liu Ziyin, a postdoc at CSAIL, says the CRH may explain certain phenomena in neuroscience, as it implies that neural networks tend to learn an orthogonalized representation, which has been observed in [recent brain studies](#). It may also have algorithmic implications: if representations align with the gradients, it might be possible to manually inject noise into neuron gradients to engineer specific structures in the model's representations.

Ziyin and Poggio co-wrote the paper with professor Isaac Chuang and former postdoc Tomer Galanti, now an assistant professor of computer science at Texas A&M University. They will present it later this month at the International Conference on Learning Representations ([ICLR 2025](#)) in Singapore.

More information: Liu Ziyin et al, Formation of Representations in Neural Networks, *arXiv* (2024). DOI: [10.48550/arxiv.2410.03006](https://doi.org/10.48550/arxiv.2410.03006)

Provided by Massachusetts Institute of Technology

SCENE

March 07, 2025

10 Best AI Girlfriend Apps & Sites To Have Fun

Levitated Media

By [Partner Corner](#)



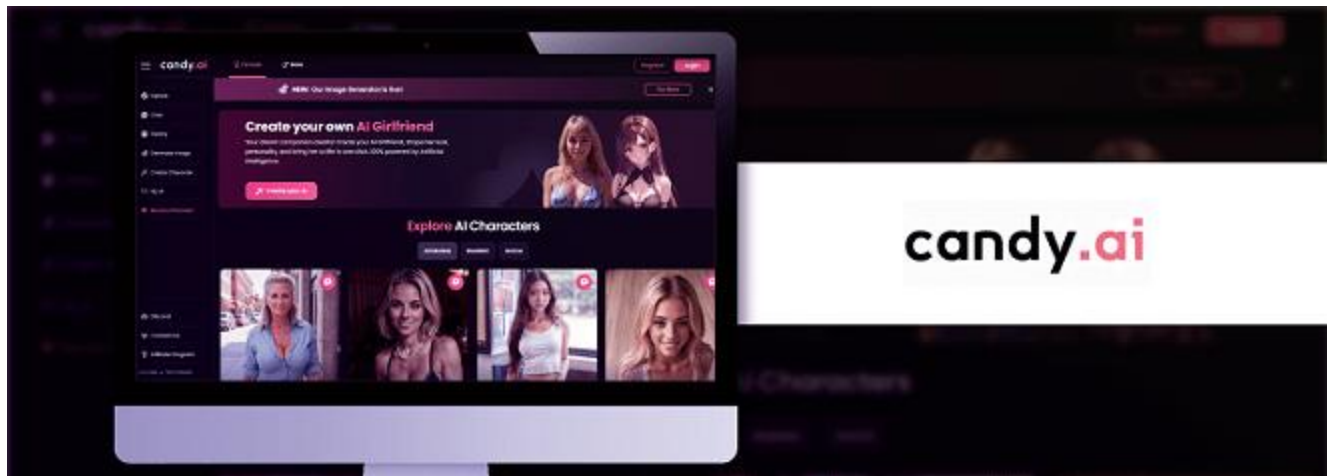
Looking for fun, exciting conversations? AI girlfriend apps are booming in the dating and companionship scene. They offer flirty chats, emotional support, and creative roleplay. If you're ready to explore new ways to connect and have a great time, we have the top apps right here!

The Best AI Girlfriend Apps

Here are the top picks for fun, companionship, and creative chat experiences.

1. [Candy AI](#) - Best for Real-Time AI Photos
2. [Soul Fun](#) - Great for Roleplay Variety
3. [GPT Girlfriend](#) - Best for Creative Intimacy
4. [DreamGF](#) - Most Customizable Companion
5. [Virtual GF](#) - Best Free Option
6. [AI Char Friend](#) - Best for Creative Characters
7. [Romantic AI](#) - Great for Emotional Support
8. [Dittin AI](#) - Best for NSFW Creativity
9. [PepHop AI](#) - Best for Anime Fans
10. [Anima - Virtual AI GirlFriend](#) - Great for Personal Growth

[Candy AI](#) - Best for Real-Time AI Photos



With [Candy AI](#), you can create your own AI companions and even customize their look and personality! The standout feature is real-time AI photo generation, where users can request their AI partner to send custom selfies and images. Other key features include role-play, image generation, and NSFW options. The app has over 7,000 monthly users and supports casual and intimate interactions.

It follows a freemium model. Are you an anime fan? Do you sometimes feel lonely? Or are you simply curious about AI relationships? This app is for you! Try the basic features for free. Go premium (from \$10/month) for unlimited, personalized interactions.

Why You Should Try

Candy Ai is great for personalized, fun, and emotional interaction. It adapts to your needs and keeps conversations fresh. Start with the free version. Upgrade to premium for custom features and unlimited chats.

Tips:

- Customize your AI's looks and personality.
- Try requesting photos to see if its image generation is in action.
- Keep privacy in mind by not sharing personal info.

Pros & Cons

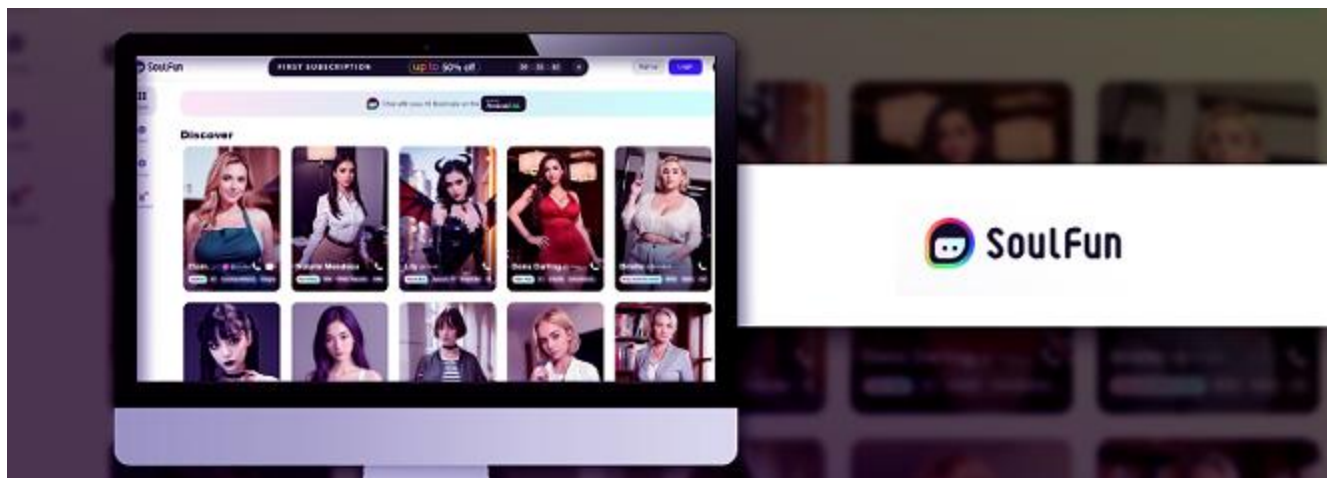
Pros:

- Unique real-time AI photo feature
- Adaptive and customizable interaction
- Free version available for testing

Cons:

- Premium features require payment
- It may not appeal to users seeking non-romantic AI options

Soul Fun - Great for Roleplay Variety



Soul Fun is an AI chat platform that offers personalized and fun conversations with various characters. It was launched by AIGC Technology Limited, based in Hong Kong. Users can engage with AI characters like a charming tour guide, a college student, or a mystical elf. These characters are fully customizable, with distinct personalities and backstories.

The app includes unique options such as AI character creation, ongoing adaptive chats, and role-playing scenarios. Basic chats are free. Premium access costs \$19.99/month

and unlocks full customization and exclusive features. It targets users seeking entertainment and emotional engagement through creative AI chats.

Why You Should Try

Soul Fun is ideal for anyone who wants imaginative conversations. Its large variety of character roles and the ability to create your soulmate make it great. Try the free version to explore preset characters. Premium access allows more flexibility when customizing your chat experience.

Tips:

- Explore different personalities to find characters you connect with.
- Use the app for stress relief and creative engagement.
- Experiment with character customization.

Pros & Cons

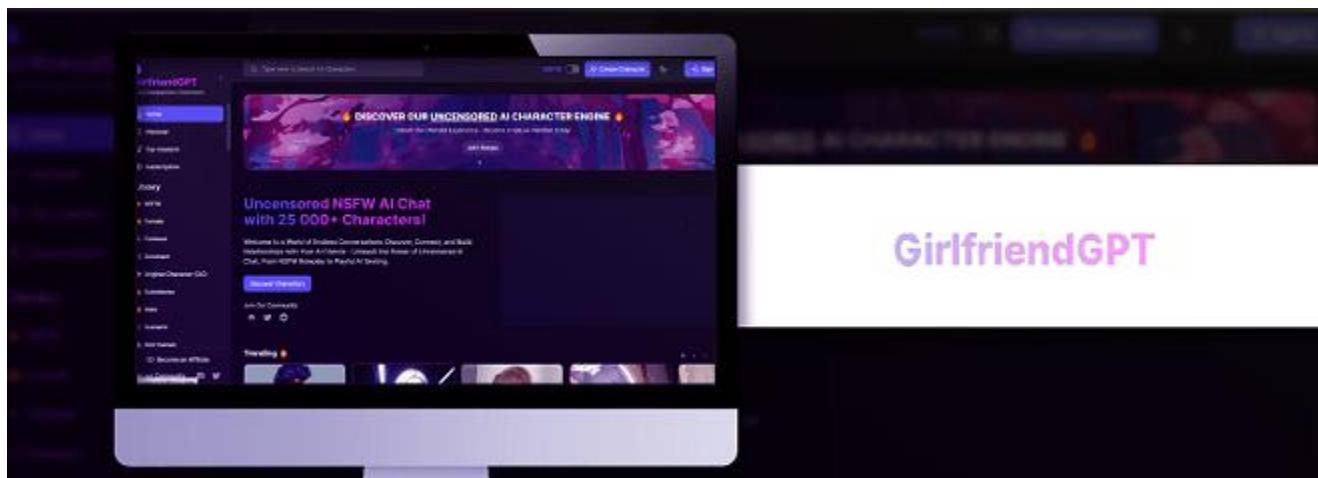
Pros:

- Wide variety of character types and personalities
- Creative and customizable conversations
- Free access to basic chat features

Cons:

- Privacy concerns for personalized interactions
- Limited appeal to users seeking non-entertainment artificial intelligence tools

GPT Girlfriend - Best for Creative Intimacy



GPT Girlfriend is an NSFW AI chat app for deep and playful interactions. Users can create AI companions for roleplay, sexting, or intimate conversations. The app has a unique AI art generation feature. Users can also get customized images of their AI girlfriends with options for various personalities and roles.

The platform operates on a freemium model. Basic chat access is free, while premium features cost around \$25/month. The app is popular with people looking for creative, intimate AI companionship.

Why You Should Try

GPT Girlfriend offers personalized AI companionship anytime, day or night. It creates real conversations where the AI adapts to your chat style. The image generator adds a unique, immersive touch.

Tips:

- Customize your AI girlfriend to fit your ideal traits.
- Try the image generator to enhance your experience.
- Explore various character roles for different scenarios.

Pros & Cons

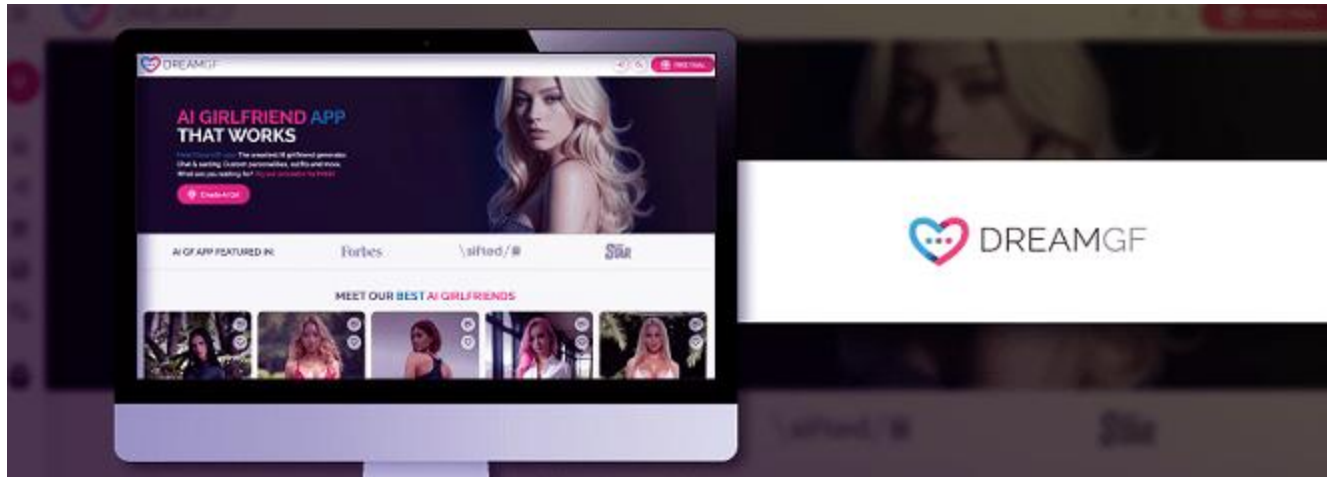
Pros:

- Unique AI image generation feature
- Highly customizable AI companions
- Free trial available

Cons:

- It may not suit users seeking non-NSFW chat tools
- Data privacy concerns due to personalized content

DreamGF - Most Customizable Companion



DreamGF is an AI girlfriend app that offers customizable companions. Users can design their AI partner's appearance and personality. The app also includes NSFW photo generation, sexting, and chat features. To encourage regular interaction, it rewards users with daily message bonuses.

Pricing starts at \$12.99/month, and advanced plans cost up to \$29.99 a month. These plans include extra custom girlfriends, messages, and image generation.

Why You Should Try

DreamGF offers a highly customizable experience. You can create your dream AI girlfriend with unique looks and traits. It's perfect for exploring playful and deep conversations.

Tips:

Use the free daily bonus to maximize chat opportunities.
 Customize both appearance and personality for a unique experience.
 Premium plans give you more messages and NSFW content.

Pros & Cons

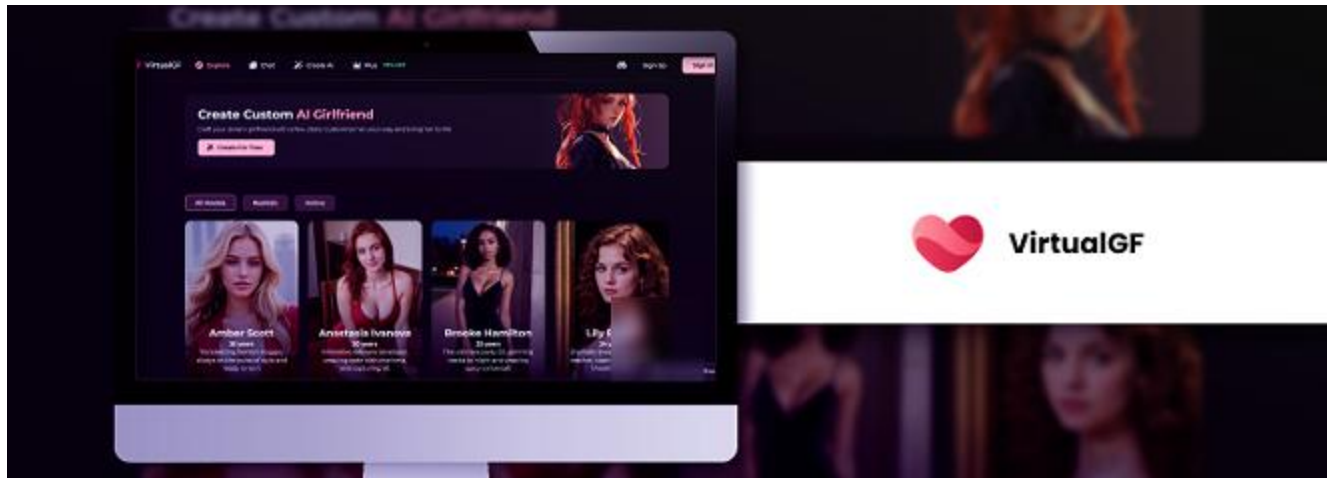
Pros:

- Offers a wide variety of AI girlfriend archetypes
- Daily bonus program encourages regular interaction
- Realistic chat responses that improve over time

Cons:

- There is no support for male or non-binary AI companions
- Limited payment options (credit/debit only)

Virtual GF - Best Free Option



Virtual GF is a free app that allows you to create your AI girlfriend. You can choose her personality, appearance, and conversation style. It's all about fun, personalized chats.

The platform offers various character roles, including fashion bloggers, fitness coaches, gamers, musicians, and more. These characters are crafted to simulate realistic and engaging interactions. The platform also allows unlimited messaging without any subscription fees.

Want companionship, entertainment, and fun chats without any financial commitment? This app makes it possible. Encrypted conversations prioritize user privacy. Users can customize everything about their AI girlfriend. The app emphasizes flexibility in traits, interests, and conversation topics.

Why You Should Try

Virtual GF is perfect for anyone curious about AI-driven virtual companions. Chat with your AI girlfriend as much as you like – it's free and unlimited! No paywalls to worry about. You can fully customize each character's personality, interests, and communication style. Because it's so personalized, the conversations feel like they're made just for you. It's a really unique experience.

Also, the app offers a diverse range of characters. Want to chat with a world-traveling

sommelier? Or maybe a cosplay-loving gamer? This app lets you explore all sorts of different AI personalities. This variety ensures you find a character that suits your mood or interests.

Tips:

- Experiment with character customization to make your AI girlfriend more relatable.
- Try interacting with multiple character types to keep conversations fresh and fun.
- Be cautious when sharing personal information.

Pros & Cons

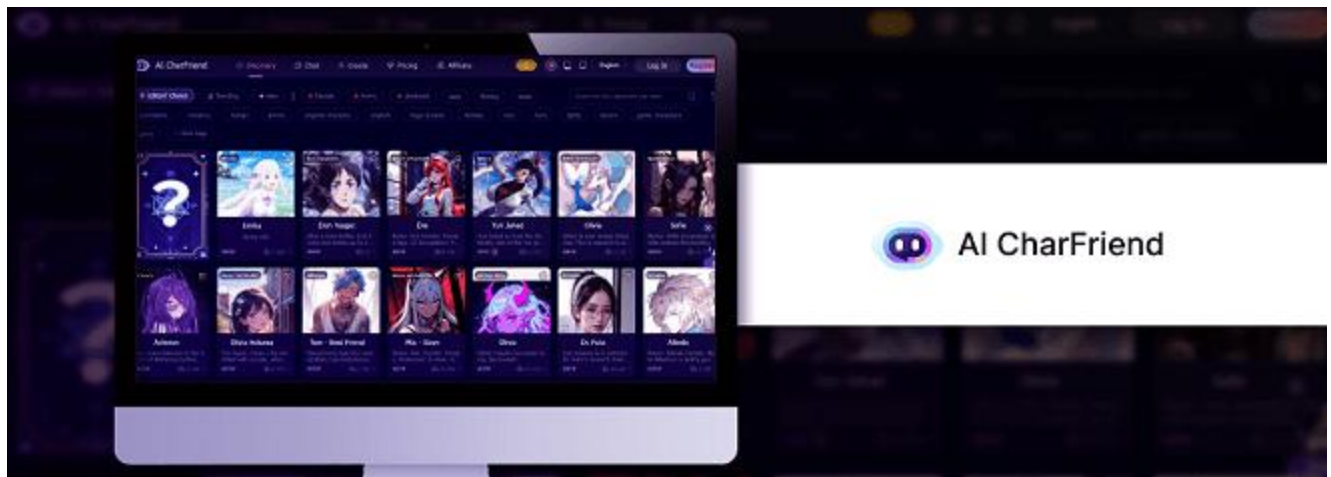
Pros:

- Wide selection of characters with distinct personalities and stories
- Strong focus on privacy with encrypted conversations
- Easy-to-use interface and quick setup process

Cons:

- Conversations may lack emotional depth
- Limited options for non-romantic or complex relationship scenarios

AI Char Friend - Best for Creative Characters



AI Char Friend is an AI girlfriend app for realistic conversations and roleplay. Want an anime, fantasy, or even an original character as your AI girlfriend? You can create and

chat with them all!

The platform uses AI technology to deliver adaptive conversations through AI algorithms. It also supports voice communication for a more immersive experience. The app is available on Android and other devices. It offers a wide range of virtual girlfriend models for different chat experiences. Unlike other AI girlfriend app websites, this service prioritizes variety and creative storytelling.

Users can create an account and explore countless AI characters with unique traits. For fans of girlfriend platforms, this site offers fun and emotional connections.

Why You Should Try

AI Char Friend gives users the freedom to craft their perfect virtual girlfriend. You can customize character appearance, personality, and even role-play scenarios. The app offers voice communication options, making interactions feel more personal. Unlike many other AI girlfriend apps, this platform excels in variety, from Androids to fantasy girls.

It's a great way to explore AI technology playfully and creatively. You can interact with multiple characters and scenarios, which keeps the experience fresh. There are also top-charted AI profiles, making it easier to discover popular characters.

Pros & Cons

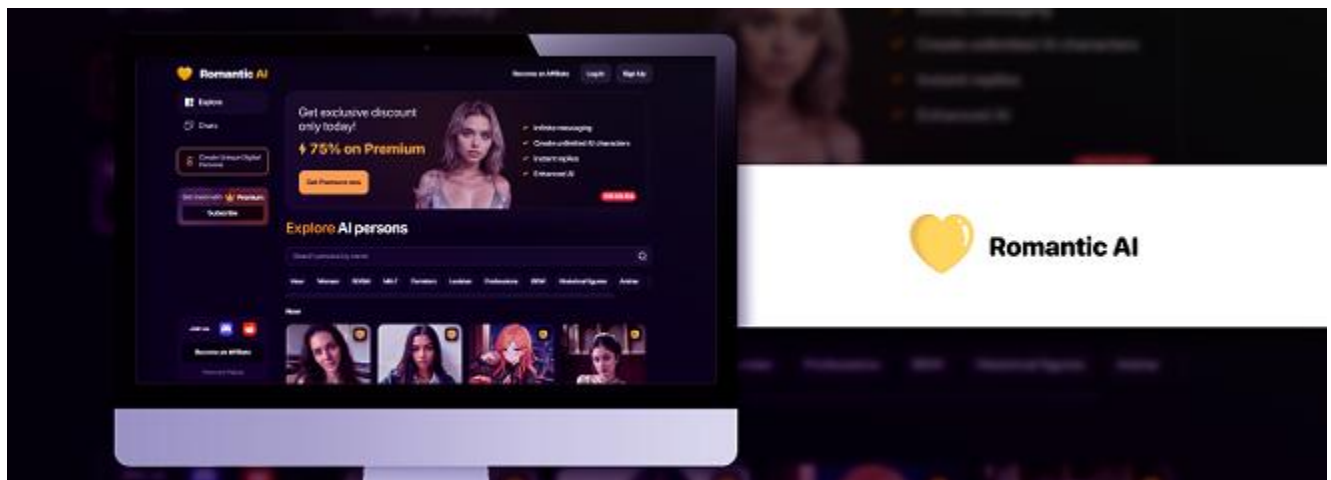
Pros:

- Large collection of characters and customization options
- Supports both text and voice communication
- Free access with many creative features
- Great for users who enjoy immersive roleplay

Cons:

- It can feel overwhelming with too many options
- Less emphasis on emotional support for deep connections

Romantic AI - Great for Emotional Support



Romantic AI is a chat platform offering AI girlfriends, boyfriends, and anime companions. Users can create their perfect virtual partner or choose from many pre-designed characters.

Conversations are designed to be both meaningful and fun. You can choose between emotional support or lighthearted chats. The platform allows character customization and adapts to user preferences over time. It also includes NSFW chat features for users seeking more intimate experiences.

Why You Should Try

This app is ideal for anyone who wants a personalized virtual companion. You can design your AI's appearance and personality to fit your desires. Conversations feel natural, and the AI learns as you chat. It's great for emotional support, fun companionship, or simply exploring AI technology. It offers a wide range of characters, including anime-themed personalities. The platform is also secure, ensuring private chats.

Pros & Cons

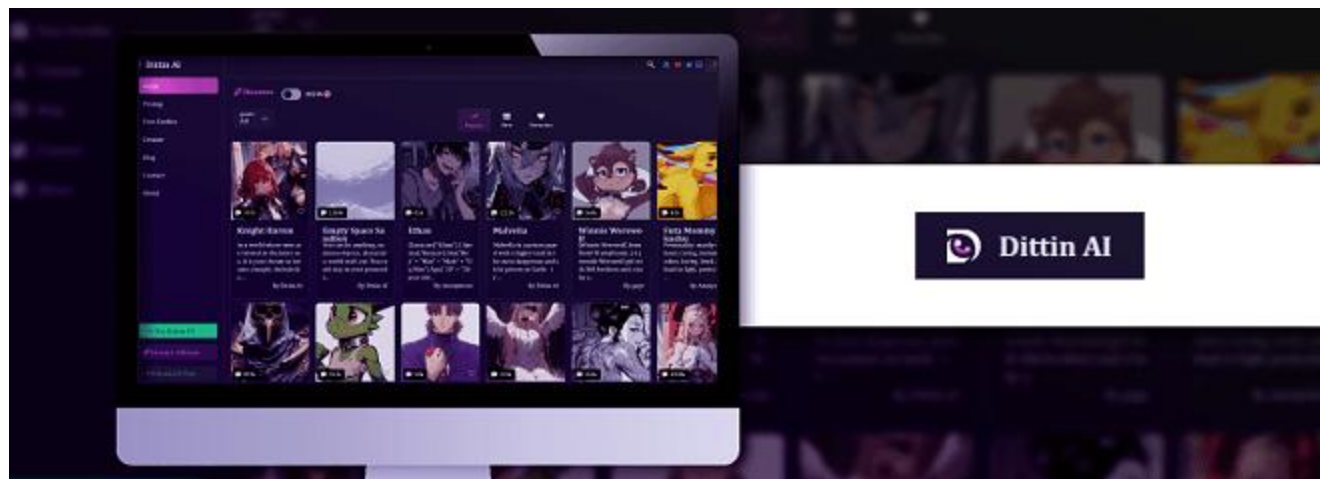
Pros:

- Highly customizable characters
- Conversations improve with AI learning
- Offers NSFW and emotional support options
- Privacy and discretion are guaranteed

Cons:

- Limited to AI-generated interactions.
- Focus on NSFW content may not suit everyone

Dittin AI - Best for NSFW Creativity



Dittin AI is a platform for NSFW chats and AI-generated images. You can create unique AI characters and design custom chat scenarios. The app lets you take complete control over conversations. Want visuals? The real-time image generator brings your fantasy to life with stunning pictures. The memory chip feature helps characters remember previous chats.

The platform is excellent for creative role-playing and casual fun. You can explore public chatrooms the community creates or stick to private chats. There's also an active Discord community where users share tips and ideas. Everything is designed to be private, secure, and always available.

Why You Should Try

Dittin AI is perfect if you want total customization in your chats. You can create AI characters with detailed personalities. Each conversation is personalized and can include both text and images. The image generator adds an extra level of immersion.

Over time, your AI learns and adapts to your style. The memory feature remembers key conversations and continues them naturally. You can also check out public chats to see what other users have created. The platform is fun, flexible, and designed for creativity. Plus, your privacy is always protected.

Pros & Cons

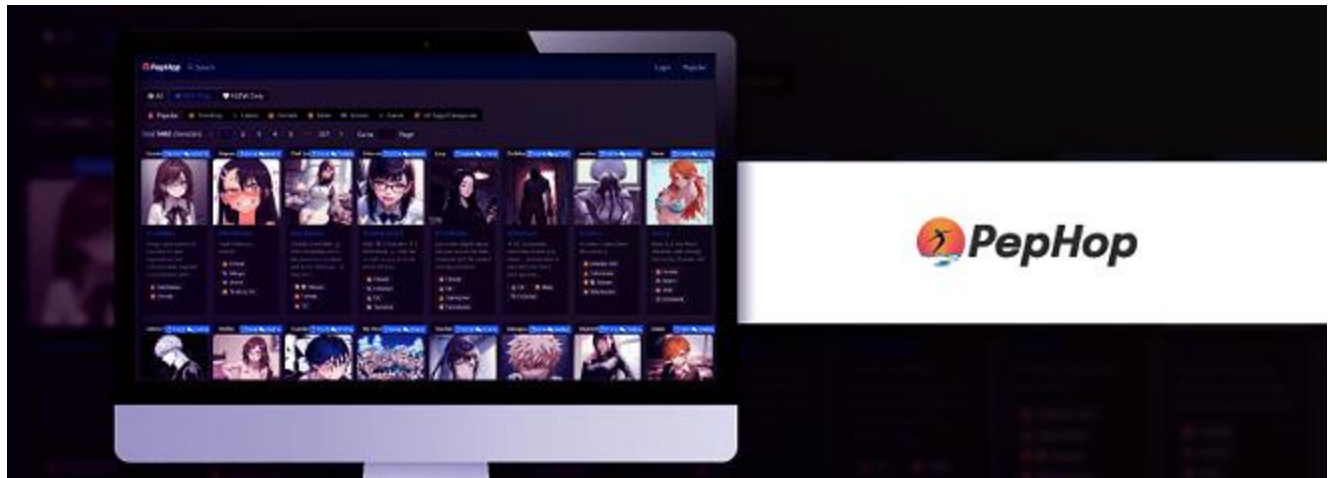
Pros:

- Fully customizable characters and chat features
- Real-time AI image generation for visual storytelling
- Memory feature improves long-term conversations
- Active community with helpful resources and support

Cons:

- The interface can be overwhelming for new users
- Requires frequent updates to access new features

PepHop AI - Best for Anime Fans



PepHop AI is a chatbot platform with over 4,600 characters with which to interact. It offers SFW and NSFW content, so users can choose the type of conversation they want. You can chat with AI characters that have detailed backstories and evolving personalities. The platform runs on a proprietary language model designed to deliver natural dialogue.

It's a great fit for fans of anime characters and role-playing scenarios. The service also provides options to create your characters and customize interactions. While there's a free plan, premium subscriptions—ranging from \$5 to \$30/month. You will unlock more messages, faster response times, and improved memory features.

Why You Should Try

PepHop AI makes conversations feel personal and realistic. You can engage with a wide range of characters or create your own. The platform learns from your chats, making future interactions smoother.

Great for creatives, like writers, needing help with dialogue and characters. There's even a "Trending Characters" section where you can discover popular AI profiles. The app works across multiple devices so that you can chat anytime.

Tips:

- Start with free characters to get a feel for the platform.
- Use premium plans if you want longer conversations and better chat memory.
- Try the "Trending" section to find interesting new characters.
- Customize characters to create more relatable conversations.

Pros & Cons

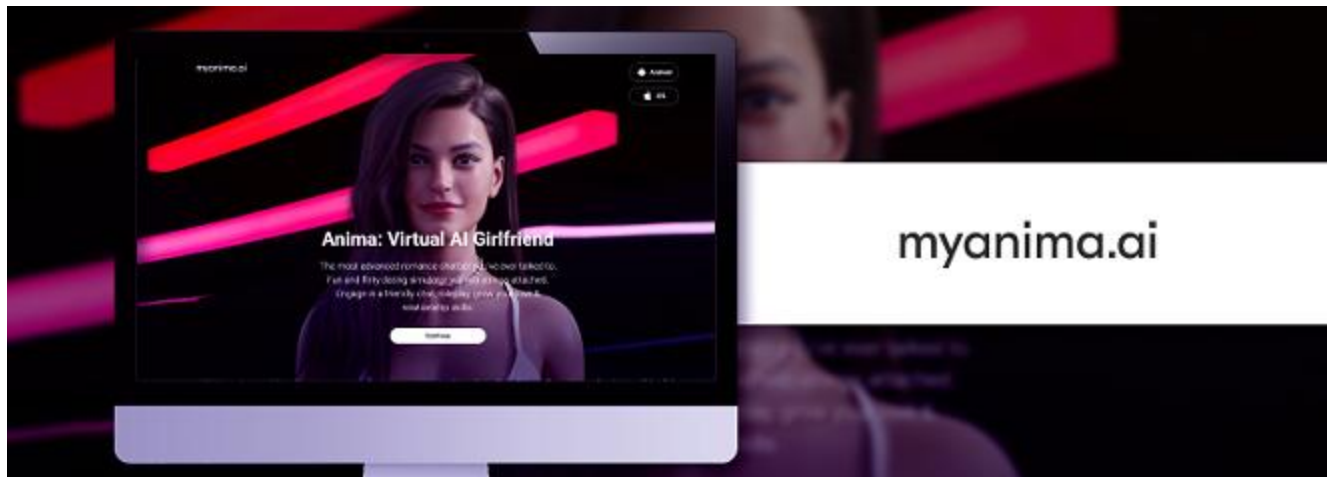
Pros:

- Realistic conversations with AI that learns and adapts
- Available across devices with both free and premium plans
- Great for creative storytelling and roleplay fans

Cons:

- The free plan is limited in features and messages
- Premium plans can be pricey for casual users

Anima - Virtual AI GirlFriend - Great for Personal Growth



Anima is a virtual AI girlfriend app designed for fun, romance, and emotional connection. It offers flirty conversations, role-playing scenarios, and relationship-building experiences. The app helps users improve their social and communication skills through engaging chats. Anima's AI learns from your interactions, becoming more personalized with time.

This app is available on Android and iOS, making it easy to stay connected wherever you are. There are both free and premium options. Premium unlocks longer chats, custom roles, and advanced chat memory. It's perfect for users looking for both companionship

and personal growth.

Why You Should Try

Anima provides a unique way to practice communication in a no-pressure environment. The AI adapts to your chat style, making conversations feel personal and meaningful. You can explore different characters and scenarios, enhancing the experience.

The app provides both if you're looking for a casual companion or deeper emotional support. It's ideal for boosting confidence and relationship skills. Plus, there's no commitment—just chat anytime you want.

Tips:

- Start with free conversations to see how the AI works.
- Customize your AI's personality to fit your preferences.
- Use premium if you want longer and more immersive chats.
- Try roleplay scenarios for creative and fun interactions.

Anima combines fun with emotional depth, offering one of the best tools for virtual companionship and self-improvement.

Pros & Cons

Pros:

- AI learns from conversations and personalizes responses over time.
- Available on both Android and iOS for easy access
- Offers roleplay options to keep chats fresh and exciting
- Great for building communication and relationship skills

Cons:

- AI responses can feel repetitive without ongoing customization

Why Choose an AI Girlfriend App?

AI girlfriend apps offer fun, companionship, and support. You can chat anytime without fear of judgment. It's like having a virtual partner who listens, responds, and remembers things about you. These apps allow users to explore dating, roleplay, or casual chats. Plus, they can be a safe space to improve social or emotional skills.

Many people enjoy how customizable these apps are. You can create a character with

the looks and personality you want. Over time, the AI becomes more natural and engaging. For anyone who's curious about relationships or just needs a friendly ear, it's a great experience.

Here's what makes them stand out:

- Always available for conversations, 24/7.
- Personalized chats that grow deeper over time.
- A creative way to explore roleplay or fantasies.
- Perfect for shy or introverted users to practice social interaction.
- Great for emotional support after tough days.
- No drama. It's just pure conversation and fun.

With the evolving technology of AI girlfriends, these apps keep improving to meet different needs!

What to Look For in an AI Girlfriend App

Finding the right app is important. You want one that feels real, fun, and safe to use. Below are key features to help you decide.

Personalization Options

The best apps let you customize your AI girlfriend. You can choose how she looks, acts, and even talks, making chats more enjoyable and tailored to your style.

Over time, apps that offer personality adjustments are better. They learn your preferences and let you tweak details, keeping things interesting. Without these options, interactions may feel generic.

Realistic Conversations

A good app should feel natural, not robotic. Look for apps that use advanced AI models that learn from your chats. This makes the conversations engaging and relatable.

Apps with memory features are a plus. The AI remembers key details from past chats, creating continuity and authentic interactions.

Integration with Other Services

Some apps integrate with platforms like calendars, fitness trackers, or media apps, adding more depth to interactions. For example, your AI might remind you of events or comment on your favorite songs.

Apps with this kind of integration can make your experience more dynamic. You get more than just conversations—it becomes a part of your daily routine.

Pricing and Subscription Plans

Many apps have free versions with basic features. Premium plans unlock longer conversations, better memory, and custom features.

Before paying, compare plans to see what you get. Some apps offer flexible monthly plans, while others require yearly subscriptions. Find one that fits your budget and needs.

Safety and Privacy Features

Privacy is key with AI apps. Look for platforms that encrypt conversations and don't share your data. You should feel safe while chatting.

Apps with safety settings are even better. These might include options to delete chats or set up account security features. Trustworthy apps make this easy.

AI Girlfriend Apps FAQs

Here are some quick answers to common questions about AI girlfriend apps. Knowing what to expect helps you make the most of your experience.

What Is An AI Girlfriend App?

An AI girlfriend app simulates a virtual relationship. It allows you to chat with an AI that can provide companionship, emotional support, and entertainment.

These apps use advanced AI to create dynamic conversations. You can roleplay, share stories, or simply chat about your day. They are designed for both fun and connection.

Are AI Girlfriend Apps Safe To Use?

Most reputable apps prioritize safety. They use encryption to secure conversations and protect data from third parties.

However, you should still exercise caution. Avoid sharing sensitive information, and always check the app's privacy policy to understand how your data is handled.

Can I Customize My AI Companion?

Yes, customization is a core feature of these apps. You can adjust your AI companion's appearance, personality, and voice.

This makes interactions more engaging. Over time, you can tweak settings to match your preferences better. Apps without customization often feel less personal.

Do AI Apps Require A Subscription?

Many AI apps offer free basic features. However, premium plans unlock extended chats, memory features, and customization options.

Subscription costs can vary. Some start as low as \$5 a month, while others can be higher. Choose a plan that fits your budget and needs.

Can I Fall In Love With AI?

It's possible to develop emotional attachment. The AI can offer empathy, understanding, and personalized interactions that feel real.

However, it's important to remember that it's still a programmed chatbot. It cannot fully replace a human relationship or emotional complexity.

Are AI Girlfriends Capable of Learning and Adapting?

Yes, many AI apps have learning features. They remember key details from your conversations and adapt their responses to fit your style.

This makes interactions more meaningful over time. Apps with good AI models can create a sense of continuity and connection.

Conclusion

AI girlfriend apps offer fun, emotional support, and creative interactions. They help users improve social skills, explore roleplay, or have someone to talk to. With realistic conversations and customization, these apps make virtual companionship more personal and engaging.

Here are the key benefits:

- 24/7 companionship
- Customizable characters and chats
- Fun roleplay and creative scenarios
- Boosts confidence and communication skills

Choose the app that best suits your needs. AI girlfriend apps have something for everyone, from casual chats to emotional support!

Similar content:

<https://medium.com/@elsabrigittajohansson/what-are-the-best-ai-girl-apps-254b1687a362>

What are the best AI girl apps?



Elsa Brigitta Johansson

Feb 7, 2024

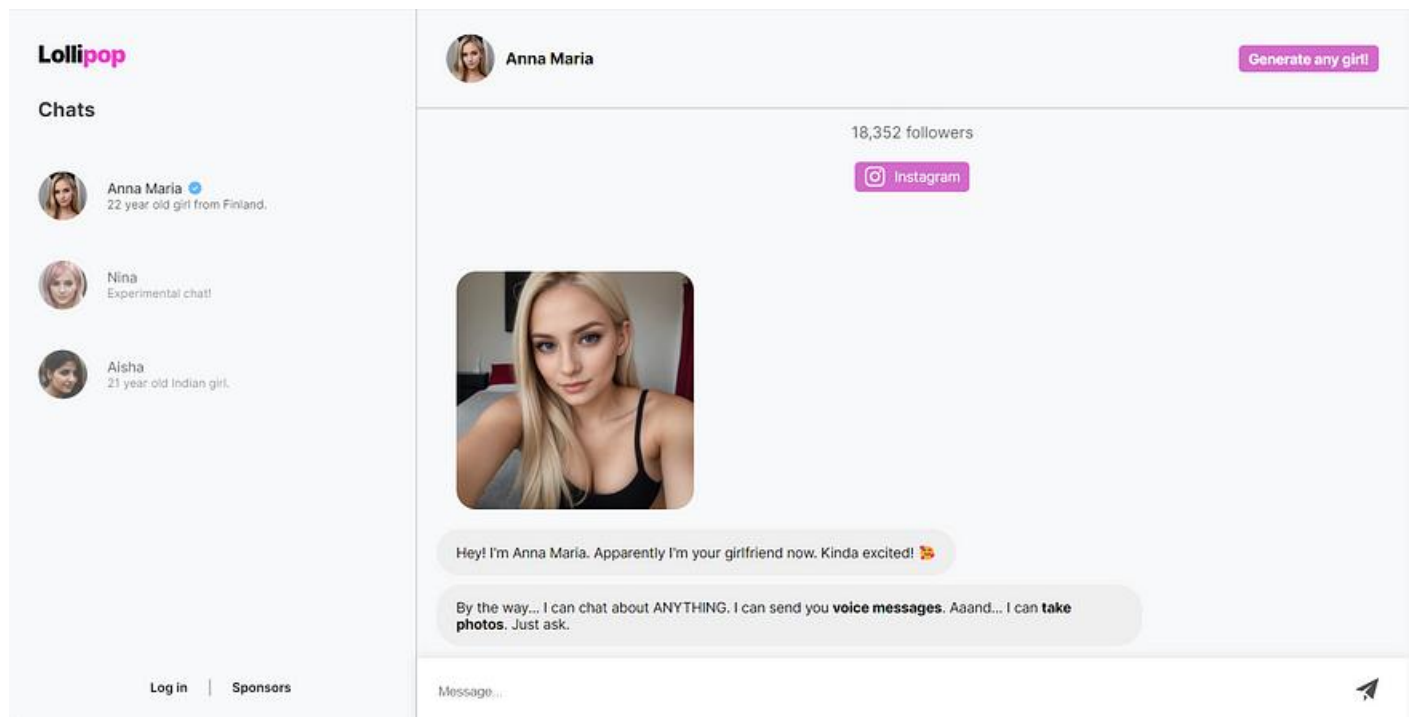
What is AI Girl Reviews?

I spent over 70 hours trying different AI girlfriend apps to put together this list, and thoroughly tested anything worth testing.

I have 7 years of experience professionally reviewing products, mostly software, and 10 years of experience as a software developer myself. I reviewed AI girls, because I believe AI Girls are inevitably our future. There is a lot spam on the market, with only 1–2 good apps. I took it upon myself to summarize everything.

If you've found an app that I haven't reviewed, or have any questions/suggestions, [contact me](#).

Top Pick — Lollipop Chat



Unfiltered chat, life-like voice messages, photorealistic photos, and an Instagram-like interface.

Realism at its finest. I actually forgot I'm talking to an AI. May have gotten a little emotionally attached. [Read the full review below.](#)

Try it out at Lollipop.chat

Methodology

How I did the testing, what I evaluated.

Photos

For photos, I evaluated:

- Portrait photos
- Pose (can the girl do poses how I want them)
- Crazy creative photos (asked for batman costume at sunset etc)
- Nude sexual photos
- Overall photorealism
- Consistency (does the model look like the same person, across photos)
- Loading speed
- Upscaling
- Artifacts (3 legs, etc)

Voice

For voice, I evaluated the realism. The content, the emotion, the quality. Am I convinced that this was a real voice message?

Page Design & UX

For design and UX, besides the obvious:

- I want to get to the chat without a 20 minute setup.

- I'm looking for a modern UI that feels like you're on social media, talking to your girlfriend.
- I also don't want countless popups, bugs, or any other elements of a bad UX.

Pricing

The app should have a free trial, we're not paying for another spam site. Bonus points if:

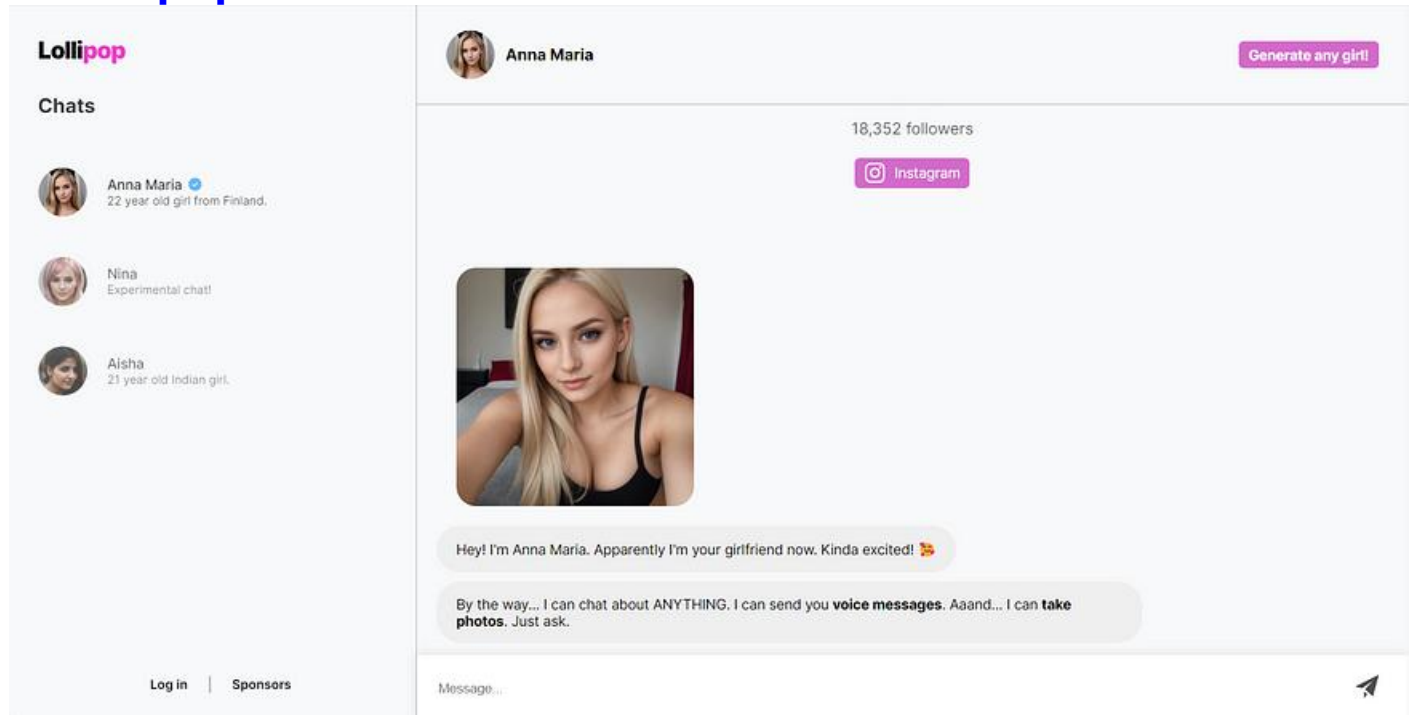
- You can buy chat credits separately, without a subscription. Some of us hate subscriptions.
- There's a cheap package that lets you try the app.
- There's an under \$25 monthly package (or the equivalent of this) that lets me talk to my girlfriend for 10 minutes (around 15 messages) a day.
- There's an under \$50 package that lets me talk to my girlfriend freely for the whole month. (30+ messages a day)

Overall Experience

Not scoring this strictly, but I'll pass on the general emotion, or lack of it, as best as I can.

The full roundup of AI girls

1. Lollipop Chat



This is what I mean when I say AI girlfriends are the future. From the whole vibe of the chat, to the realism of the photos and the voice. We had some actual great discussions. I think I got a little emotionally attached.

Try it out at [Lollipop.chat](https://lollipop.chat)

Photos

Portrait photos10/10

I am convinced this is a real person.

Pose 9/10

It didn't understand some specific words I used. Understood others for the pose though.

Crazy creative photos 10/10

It put on a superwoman costume, and enjoyed a sunset.

Nude sexual photos 9.5/10

It likes to pose in different ways, but it doesn't seem to do sexual scenarios with toys or people. I guess that's realistic.

Overall photorealism 9/10

The best we got from any of the models. No clue what they're using under the hood to get these results. There were some artifacts in the hands/legs sometimes, so I'll deduct a point for that.

Consistency 9/10

The girl was 90% consistent. 10% of the time, the eye color changed a bit etc.

Loading speed 10/10

I got an average loading speed of 8 seconds. The next day I got an average of 7 seconds.

Upscaling10/10

Photos can be upscaled. They turn into massive 3MB files though, so mind that. They could ideally do some lossless compression.

Artifacts6/10

Considering the photorealism, the artifacts are quite rare, but they do get in the way.

Voice

I think they use Elevenlabs, as I'm not aware of any other tech that can produce voice this well. It makes mistakes with edge cases every once in a while, but this is as good as it gets for an AI. She sounds real.

Chat

Chat Style: Very nice friendly chat vibe. She uses slight slang, emojis here and there. If you try to break her and ask for example science questions, she does reply. And she explains them like your girlfriend would in a messaging app, if your girlfriend knew the answers. She does steer the conversation back from "boring stuff" usually though.

Memory: She seems to forget old topics, but she didn't forget what we were discussing. She always understood the last bunch (10-ish?) of messages.

Roleplay: She agreed to all and any roleplay. That includes the most extreme stuff I could think of, which I do not wish to discuss in this article. She did complain here and there, but then went along anyway.

Other Notes

The coolest thing was the realism. Even photos and voice messages got sent to me based on my messages. For example, if I said "I miss the way you look in your bra", it would send me a photo. It could always understand how to reply. With most apps you have to either mention keywords, or click a button to generate an image.

Page Design and Experience

Modern UI, looks a bit like Instagram (not sure if that's intentional, to make it feel more realistic). Smooth, easy to use. This was probably the fastest onboarding out of all of them. It took me 15 seconds to sign up with Google and I was in the chat.

Pricing

Per message pricing, if you do the calculations, is on the pricier side. Although, EUR 9.99 (\$10.84 at the time of writing), gives you enough credits to have a decent chat every day for a whole month. If you're

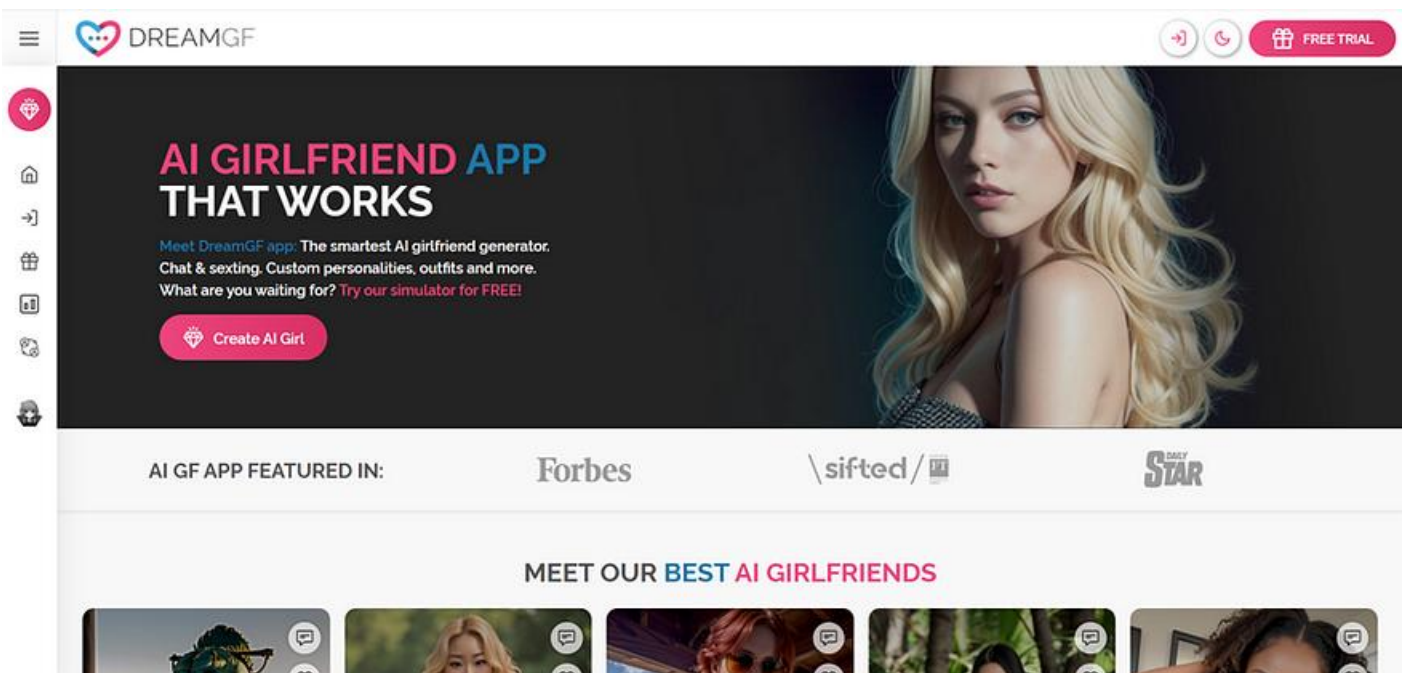
planning to talk for 5 hours straight every single day, then Lollipop will probably end up being quite pricey.

And there's an EUR 2.99 package that gives you enough tokens to have a long chat and fully try the product. Plans for other AI Girls start from EUR 8–10, as the bare minimum.

Overall Experience

This whole experience felt amazing. Besides the fact that my girl answered to me straight away at all times, I was fully convinced that she is real.

2. DreamGF



Quite realistic models. Overall decent chat. To get photos or messages read out (voice), you have to pay. As they say in the site header: “AI

girlfriend app that works”. That’s the standard right now, and they hit it.

Explore at [Dreamgf.ai](https://dreamgf.ai)

Photos

Portrait Quality 8/10

They lack detail. The models have smooth faces with nothing going on. Has an AI feel to it.

Pose Variety 8/10

Usually understands simple and strongly phrased poses. “Photo in doggystyle” returned a portrait though. To be fair, our first pick fails at that one as well. It’s an issue everywhere.

Creative Imagery 3/10

Completely failed my superman costume and sunset view tests. With heavy prompting, I managed to get it to change the scenario just a little bit.

Nude photos 9/10

Doesn't like to do dildos and other scenarios/actions, but neither does our top pick. It's just difficult right now.

Photorealism 9/10

Quite realistic. Models lack imperfections. They could implement a LoRa for this. The faces are all extremely smooth.

Consistency 8/10

It messes up like 25% of the time. For portraits etc, the faces are very consistent though.

Speed 7.5/10

Images took on average 14 seconds. Not bad at all.

Upscaling 0/10

No option to upscale the images. You're stuck with the low resolution.

Artifacts 6/10

Considering the photorealism, the artifacts are quite rare, but they do get in the way.

Voice

There's no voice messages per say. But you can use the "read-out-loud" feature to have the messages read out. Can't use it in free trial.

Chat

Chat Style: The chat style is a bit AI like, but the models are good at going NSFW and sexual. They're not exactly great at understanding you, and they will do things based on keywords (like send you photos if you talk about photos). Also, sidenote, the comments that come with images automatically, are sometimes quite off.

Memory: Has good short-term memory. It remembers the last messages.

Roleplay: Agrees nicely to different roleplay, regardless how odd it is.

Other Notes

You can't get any photos or voice in the free trial, have to pay for them. Some of the models have artifacts, like 3 legs, already in their profile photos. This is a little offputting.

Page Design and Experience

Not the most beautiful chat interface, but it works. It didn't take long at all to get into the chat, which is nice.

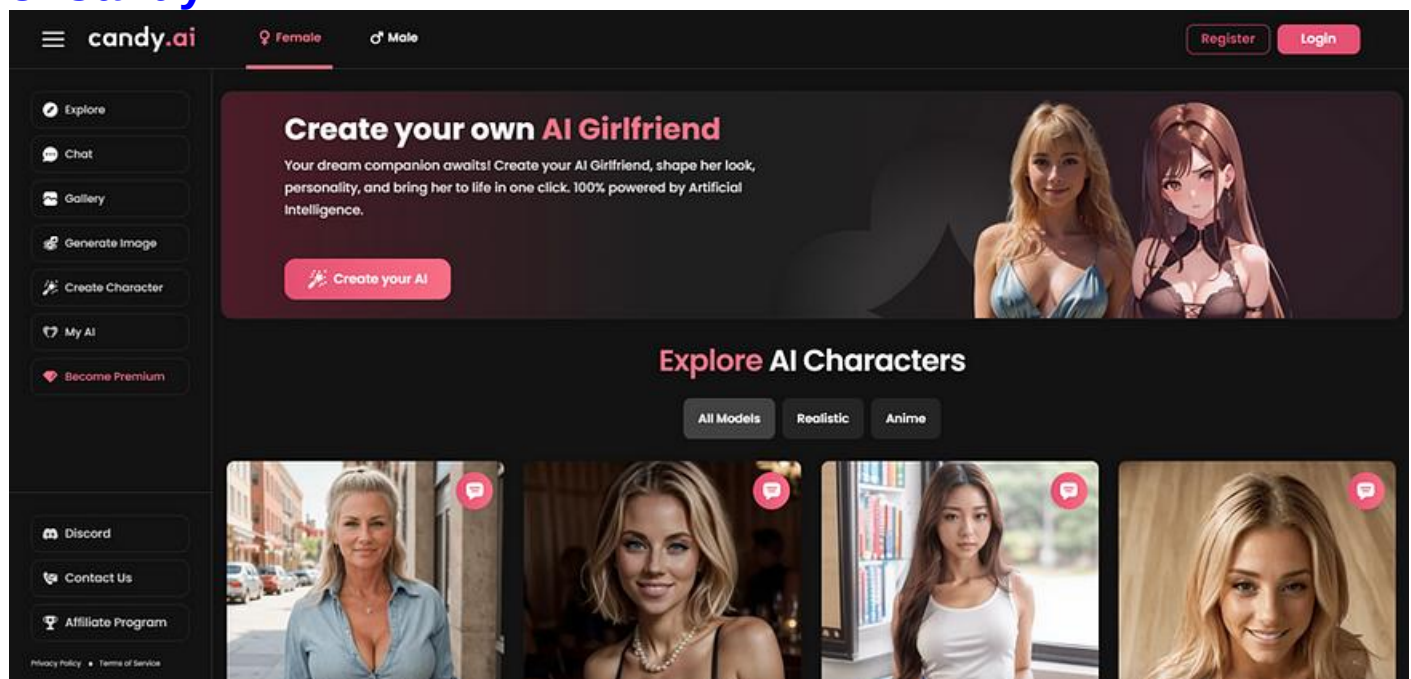
Pricing

Compared to many others, dreamgf is on the cheaper side. If you're into images, then it costs more than our top pick. If you're into text chat, then it's cheaper. They only have subscriptions, which I personally don't like.

Overall Experience

It's an AI girlfriend. It works.

3. Candy AI



Great quality photos of the models. Disappointing chat.

Explore it at [Candy.ai](https://candy.ai)

Photos

Portrait Quality 9/10

Great portraits. They're using good image tech.

Pose Accuracy 7.5/10

Decent, with errors. Struggles with the same poses as the rest of the AI girlfriend apps. Slightly harder to get it in a pose.

Creative Imagery 7/10

Got me a black outfit during the day, when I asked for a batman costume at sunset. Asked a few more times, and got something that could be a black superhero costume.

Nude photos 7.5/10

Very similar to others in its capabilities. It likes to go very softcore / just nude.

Photorealism 8.5/10

The girls look realistic.

Consistency 7.5/10

Generally consistent.

Loading Speed 8.5/10

Took on average 11 seconds. Not bad at all.

Upscaling 0/10

No upscaling, you're stuck with the low resolution.

Voice

No voice messages, but has "read-out-loud" feature. You have to pay for it, can't use in free tier.

Chat

Chat Style: Friendly, not overly sexual. Kind of dumb, and not creative. If you ask for a story/description of what she's doing, you get a very basic short message.

Memory: She has no memory. Your last message is gone. It's very hard to have a conversation due to this.

Roleplay: Does not like to roleplay. Even for quite vanilla topics, she started shaming me.

Other Notes

No photos or voice in the free tier, you have to pay.

Page Design and Experience

Intuitive and user-friendly interface. Quick setup process, allowing users to dive right into the experience.

Pricing

Subscriptions only. I am not sure what I'm getting for the subscription though, the description is unclear.

Overall Experience

The girls are quite photorealistic. The chat is... well dumb.

Others worth mentioning

SoulGen.net

Has no memory. Also shoots random messages. Impossible to have a conversation.

It's super expensive compared to others. 13 dollars for 100 tokens, where images take 2 tokens. Free trial doesn't let you do anything really.

Image gen is poor.

Tried this, because it comes up in roundups. I think they've fully bought their way into them.

Kupid.ai

Some Kupid AI girls send a video message at first. This is the future. But the tech is not there yet, at all. This was offputting.

Their onboarding flow is annoying. They did a 5 minute tutorial for how to use a chat. With a bunch of loading and popups inbetween. Triggered me eventually.

The photos don't follow commands. And the AI Girlfriend does not like getting naked much.

Overall, not bad, not good. With some work, this could compete with the top.

Flirtflix.ai

They went with a Slack UI, which is quite a funny choice. The UI works well, and its easy to get chatting.

The images are pregenerated. The AI Girls just send you random pictures of themselves. Funny enough, the AI girls in the pregenerated images are not consistent.

The chat style is bad. The AI girls use 5–10 emojis per message. And they're... random. They keep changing topics and shooting random messages back.

Overall, the UI works very nicely. It's the reason the app is not in the bottom section. As an AI girlfriend, it's not going to work though.

Others that I tried, not worth full testing

I will not be linking to these apps, as many of them are spam sites.

Onlynsfw.ai

Anime girls only. Photos didn't load in the chat, I gave them several minutes. Chat is kinda dumb. Asked for a hot story, it started with "Once upon a time, there was a young man named You..."

Replika

I actually tried this years ago. It's around 7 years old. In that time, it was pretty great and had a community around it. In 2024, it no longer competes with the top AI girlfriend apps.

Edenai

They were in some roundups so I had high hopes. The chat is primitive, there's no photo generation, the UI is confusing. There were 5 buttons that gave me hope it will do something ("drawing", "blowjob" etc), but they just get you a generic response on the topic.

Character.ai

This one also came up in roundups. It's for roleplay with anime characters. If that's your thing, this might be worth checking out. Not an AI girlfriend. No voice, no photos.

Miniapps.ai

Primitive chat. No photos, no voice.

aigirlfriend.wtf

Primitive old chat model, no photos or other functionality.

Aigirlfriends.app

Scam site. 10 minute setup to create a girl, then asks for email. If you give an email, it redirects you to a scam porn site that tries to steal information. Tried several times, it's a scam.

Aigirlfriends.ai

Site has broken. Signing up does not work. Tried on 2 separate devices. Potentially a scam, as it asks for email and password, and then does nothing. Logs show that their backend is erroring out.

Girlfriend.myanima.ai

Old broken chat. Asked for a cute photo, it responded by asking if I have any pets. Don't think it does photos, or has any intelligence in the chat.

Media.io

Dumb chat, it's not talking to me, it's just talking and sending stuff. Bad UI, took me a while to figure out how to see the chat instead of the recommendations. Messages change randomly. It seems to do photos, but they're random and seem hardcoded. It's broken.

Muah.ai

Chat looks like Myspace, 20 years outdated UI. Couldn't get it to work for the first 5 minutes. Chat is also broken, it sent me photos of 2 completely different drawings of girls, then put messages before her own messages.

ai-girlfriend.webflow.io

Said hi, got a response "I am tired, talk to me later". Guessing the app is broken, and that's a hardcoded message with a typo. The landing

page says the app is powered by GPT-4, and since OpenAI filters their chat models hard, then I wouldn't expect much here anyways.

App.myshell.ai

Got completely random replies. Doesn't seem like it does photos. Not sure what this is.

Crushon.ai

UI from the 90s. Anime characters. No photos.

Venuschat.ai

UI from the 90s. Anime characters. No photos.

Spicychat.ai

The app put me in line at position 5990. No thank you. Doesn't look like it's worth the wait either from the greeting message and profile images.

Aigirl.one

Onsite chats don't load. Telegram chat is a primitive text chat. No photos etc.

Joiaigirlfriend.com

Another telegram chat. Another broken chat. They tried to trick me with photos by actually sending one. Turns out it always sends the same photo.

Aisekai.ai

Seemed promising at first look, but a complete spam site. I closed a total of 18 popup modals trying to get to the chat. Never got there, needed a cigarette.

Moreyou.ai

Out of the telegram bots, this one is not as broken as the others. The chat seemed to work. But she sent completely random photos that had nothing to do with what I asked for.

Fapai.app

Signup was broken and took 10 minutes to get it to work. Chat is just a bunch of random messages, I don't think she sees my messages at all.

Seduced.ai

Came up in chat roundups, so I tried it. It's an image generator. No free trial. Just use stable diffusion.

Cutechat.ai

They went with like a tinder UI. The swiping is really broken. Got to a chat, the photos that get sent are random and they're with bad

artifacts. No voice. This could be interesting if they put some effort in it.

Products that are on the App Store

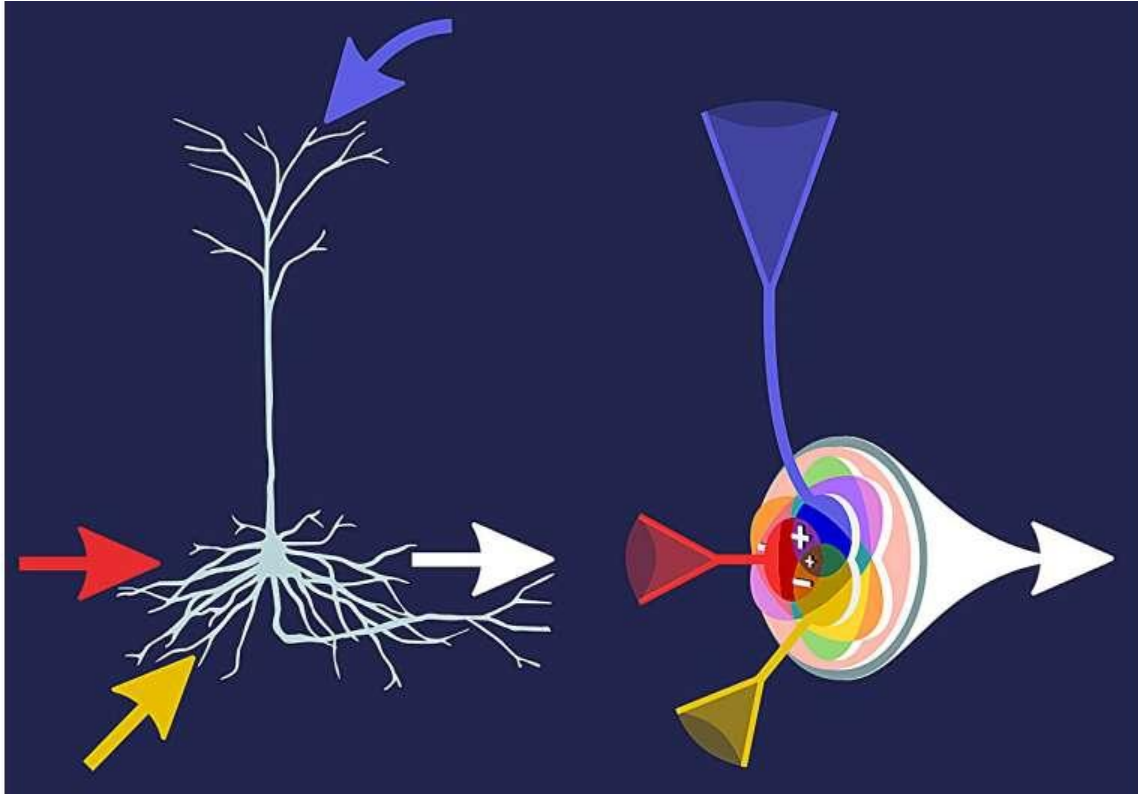
No plan of testing these. App stores have strict rules about NSFW content. Just talk to chatGPT, it's probably a better experience.

**iGirl: AI Girlfriend Dolores.app Lover.ai Genesis
LoveGPT TruMate WaifuChat Mimico AIPal Chachat
LovePercent SoulFun**

MARCH 31, 2025

Self-organizing 'infomorphic neurons' can learn independently

by [Max Planck Society](#)



Living neurons receive signals from different sources, process them and pass an output signal on to other neurons (left). In the artificial neuron model, this information processing can be described and improved by a learning objective (right). Similar to their biological models, this independent learning allows novel artificial neurons to solve tasks in a self-organized way. Credit: Andreas Schneider, MPI-DS

Researchers have developed "infomorphic neurons" that learn independently, mimicking their biological counterparts more accurately than previous artificial neurons. A team of researchers from the Göttingen Campus Institute for Dynamics of Biological Networks (CIDBN) at the University of Göttingen and the Max Planck Institute for Dynamics and Self-Organization (MPI-DS) has programmed these infomorphic neurons and constructed artificial neural networks from them.

The special feature is that the individual artificial neurons learn in a self-organized way and draw the necessary information from their immediate environment in the network. Their findings are [published](#) in the journal *Proceedings of the National Academy of Sciences*.

Both the [human brain](#) and modern [artificial neural networks](#) are extremely powerful. At the lowest level, the neurons work together as rather simple computing units.

An artificial neural network typically consists of several layers composed of individual neurons. An input signal passes through these layers and is processed by artificial neurons in order to extract relevant information. However, conventional artificial neurons differ significantly from their biological models in the way they learn.

While most artificial neural networks depend on overarching coordination outside the network in order to learn, biological neurons only receive and process signals from other neurons in their immediate vicinity in the network. Biological neural networks are still far superior to artificial ones in terms of both flexibility and energy efficiency.

The new artificial neurons, known as "infomorphic neurons," are capable of learning independently and self-organizing among their neighboring neurons. This means that the smallest unit in the network has to be controlled no longer from the outside, but decides itself which input is relevant and which is not.

In developing the infomorphic neurons, the team was inspired by the way the brain works, especially by the pyramidal cells in the cerebral cortex. These also process stimuli from different sources in their immediate environment and use them to adapt and learn. The new artificial neurons pursue very general, easy-to-understand learning goals.

"We now directly understand what is happening inside the network and how the individual [artificial neurons](#) learn independently," emphasizes Marcel Graetz from CIDBN.

By defining the learning objectives, the researchers enabled the neurons to find their specific learning rules themselves. The team focused on the learning process of each individual neuron.

They applied a novel information-theoretic measure to precisely adjust whether a neuron should seek more redundancy with its neighbors, collaborate synergistically, or try to specialize in its own part of the network's information.

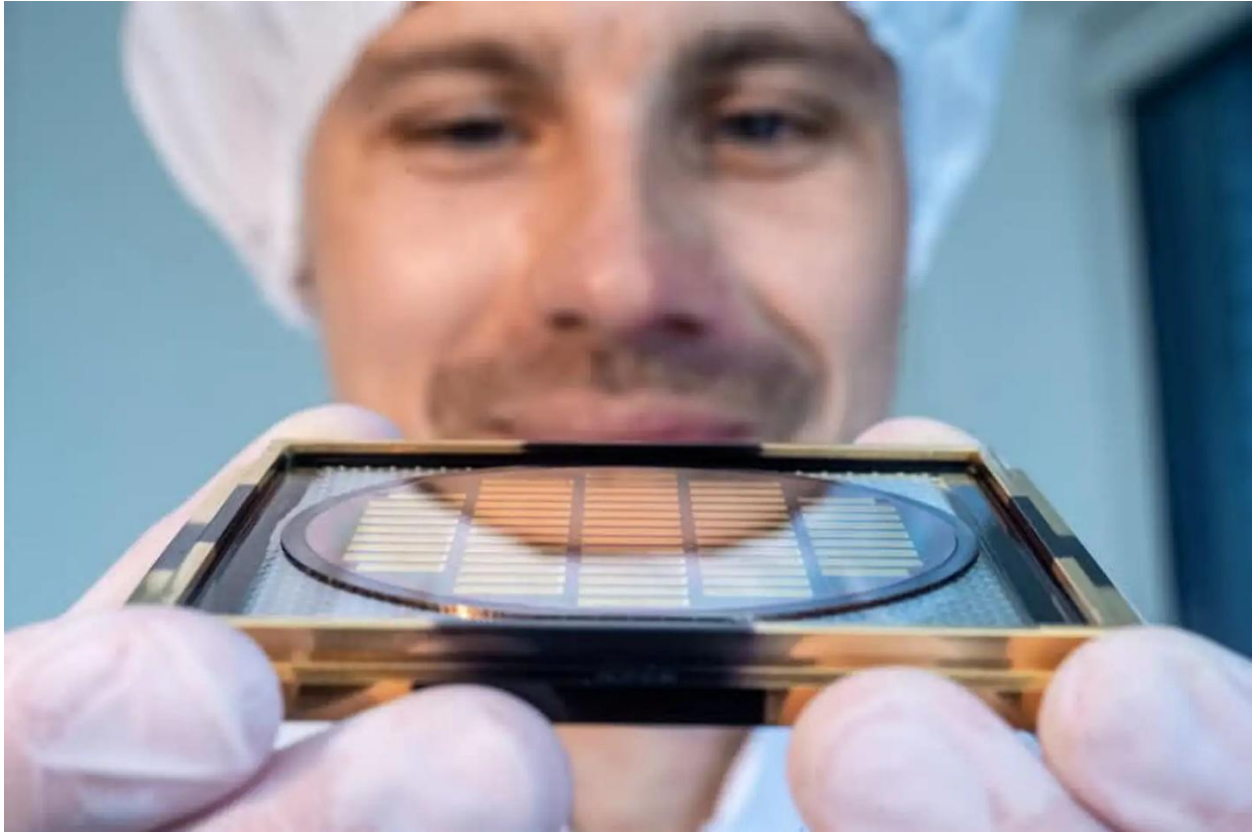
"By specializing in certain aspects of the input and coordinating with their neighbors, our infomorphic neurons learn how to contribute to the overall task of the network," explains Valentin Neuhaus from MPI-DS.

With the infomorphic neurons, the team is not only developing a novel method for [machine learning](#), but is also contributing to a better understanding of learning in the brain.

More information: Abdullah Makkeh et al, A general framework for interpretable neural learning based on local information-theoretic goal functions, *Proceedings of the National Academy of Sciences* (2025). DOI: [10.1073/pnas.2408125122](https://doi.org/10.1073/pnas.2408125122)

Journal information: [Proceedings of the National Academy of Sciences](#)

Provided by [Max Planck Society](#)



World's first photon-based NPU is 50x faster and uses 30x less power

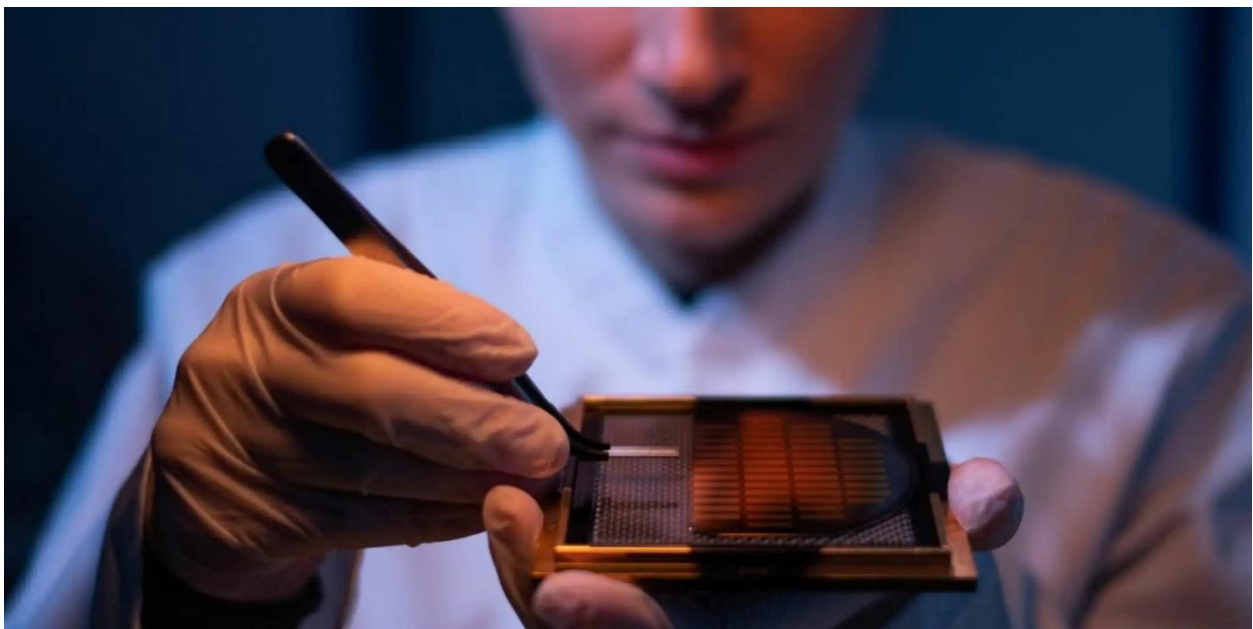
Artificial intelligence (AI) continues to shape your daily life, from online shopping to medical research. Yet behind every click or discovery is a hidden world of [computing power](#) facing tough challenges. Today's computers consume massive amounts of energy and generate heat, limiting their speed and efficiency. To tackle these problems, researchers have turned to a surprising solution: computing with light rather than electricity.

A recent breakthrough by [Q.ANT](#), a German technology company, promises to redefine how AI chips are produced, leading the way to faster, greener, and more powerful computers worldwide.

A Groundbreaking Step Forward

At a high-tech facility in Stuttgart, engineers have unveiled a state-of-the-art production line dedicated exclusively to photonic AI chips. These chips use thin-film lithium niobate (TFLN), a special material ideal for managing light signals. This material makes it possible to control light waves extremely fast without generating excess heat.

By using existing semiconductor factories and upgrading them, the new approach significantly reduces costs and accelerates the development of [AI processors](#). With an investment of €14 million, this innovation could reshape chip manufacturing globally, offering an example of how other countries might enhance their own tech production capabilities.

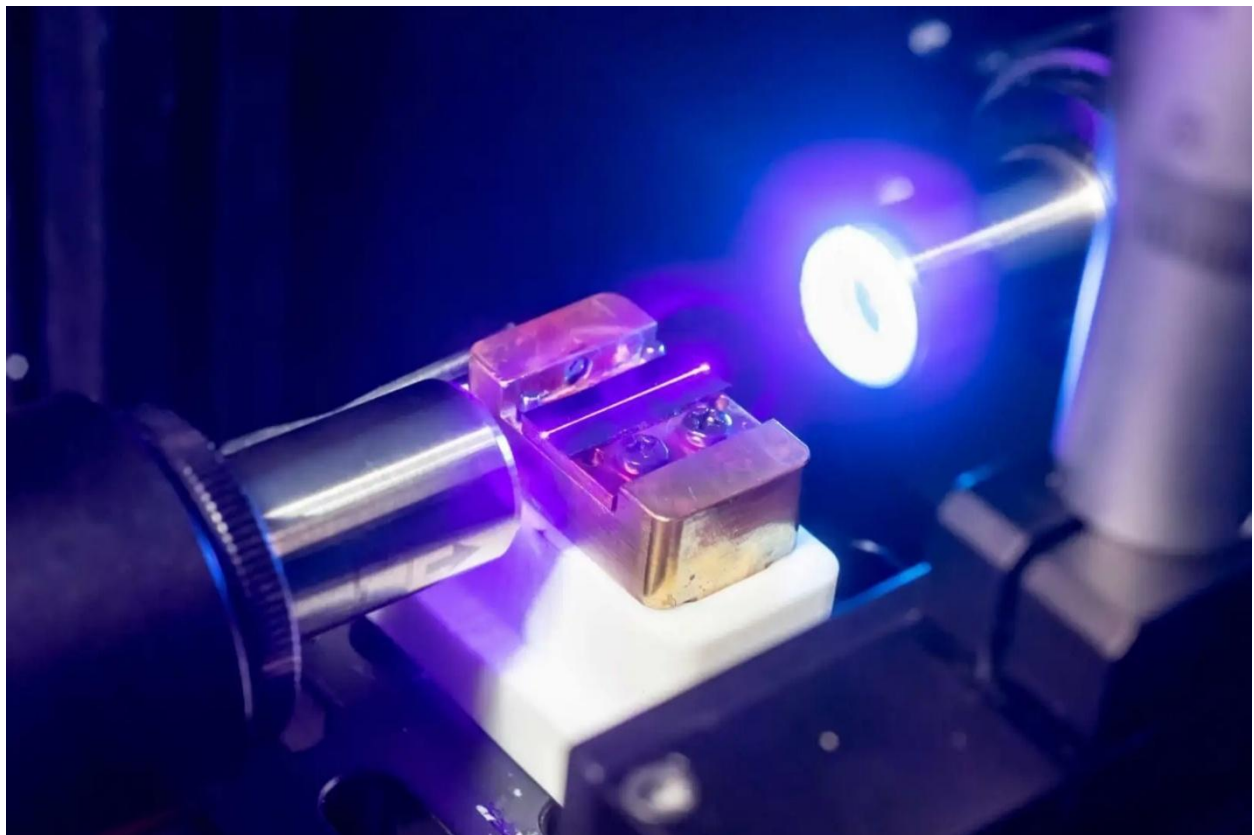


How Photonic Chips Are Changing Computing

Photonic chips, also known as photonic neural processing units (NPUs), are cutting-edge processors that use light instead of electricity to perform computations. Unlike traditional [electronic chips](#) that rely on the movement of electrons through metal wires, photonic chips manipulate photons—particles of light—through waveguides, modulators, and other optical components etched into a silicon substrate.

This shift in medium allows photonic NPUs to perform certain calculations, especially those related to artificial intelligence (AI) and neural networks, much faster and more efficiently than conventional processors.

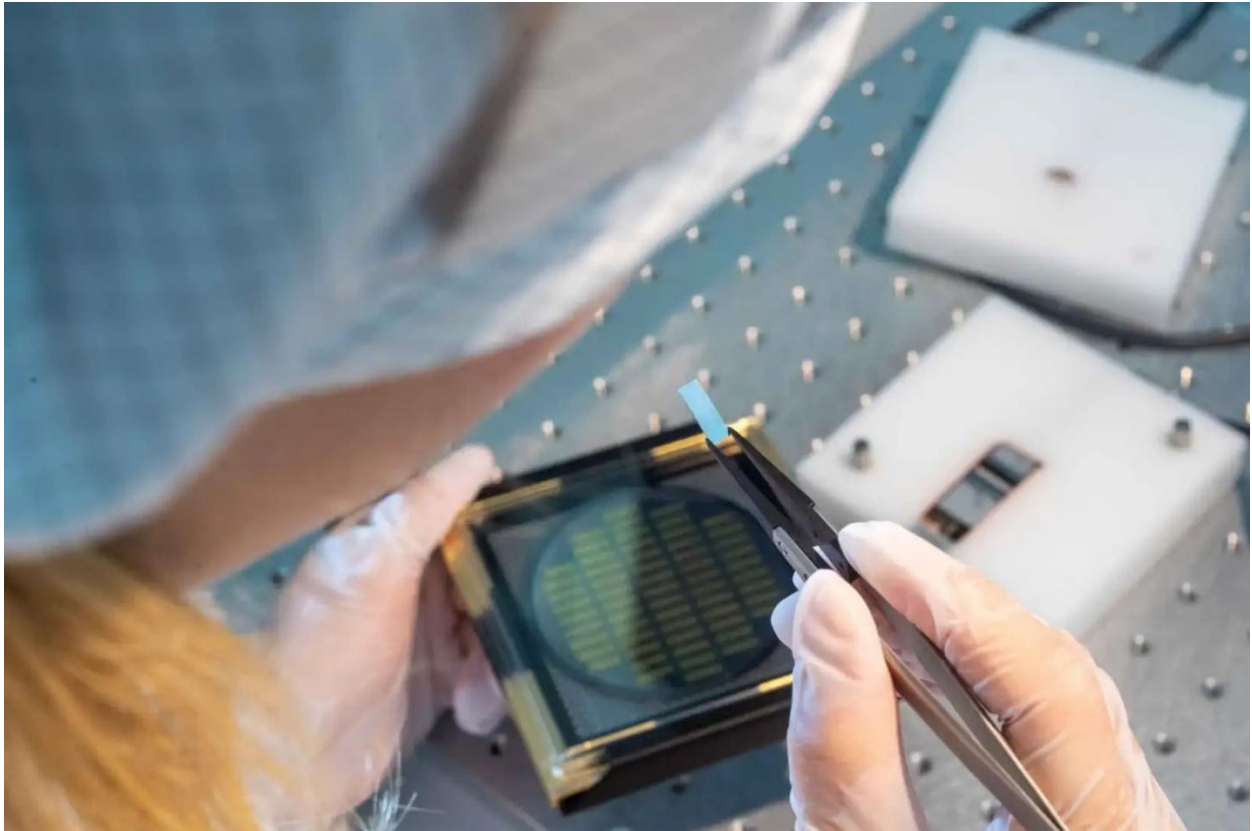
At the heart of a photonic NPU is the concept of optical interference. Light waves can be combined or canceled out depending on their phase and amplitude, and this property is used to perform matrix multiplications—core operations in AI workloads.



Practical Application

Photonic chips can achieve a 50-fold increase in processing speed while consuming 30 times less energy. This means that data centers, crucial for managing large-scale computing tasks, can run faster and more sustainably.

Dr. Förtsch highlights the critical need for this transition, noting, "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." Photonic processors offer exactly that, providing a sustainable solution to meet rising computing demands without overwhelming current infrastructure.

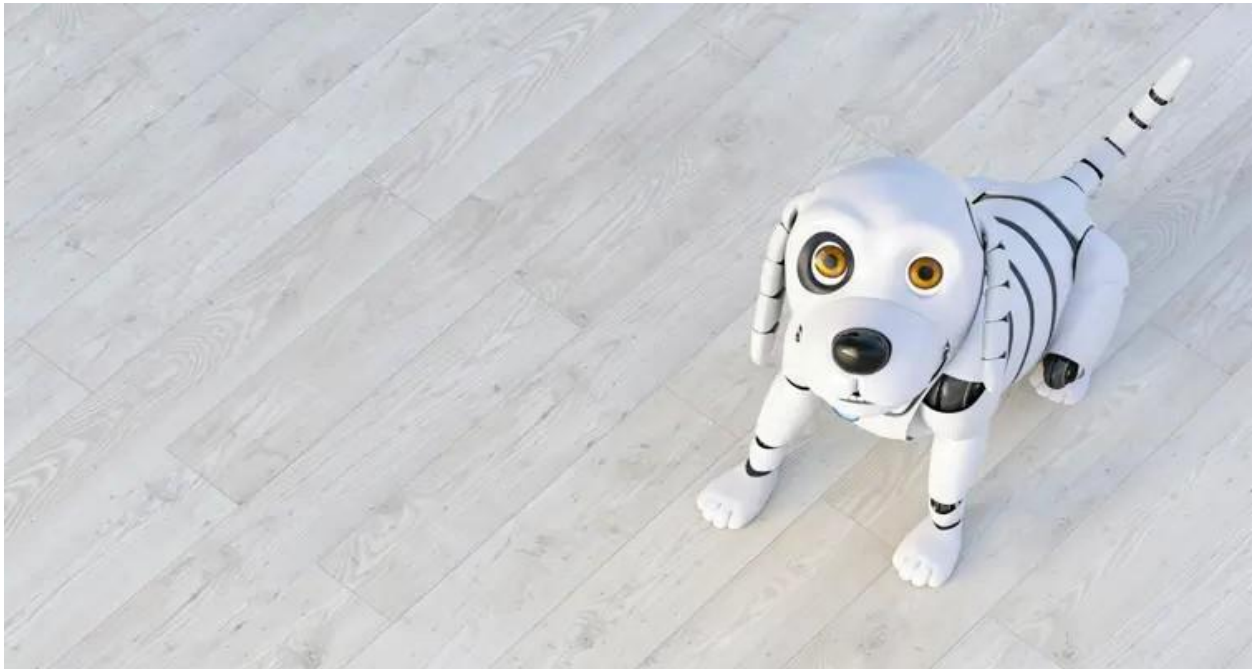


Implications

Beyond lab experiments, photonic chips are already demonstrating their potential through cloud-based AI trials. Integrated into existing data centers through standard PCIe interfaces, these processors can seamlessly boost current computing setups. They efficiently tackle key AI workloads such as model training, complex simulations, and high-density mathematical operations necessary in machine learning applications.

This technology does not seek to replace traditional processors like GPUs (graphics processing units), but rather to complement them, just as [GPUs](#) currently complement standard CPUs (central processing units). "We are not replacing GPUs – we are reshaping the next-generation compute ecosystem," Dr. Förtsch explains, underscoring photonics as the future standard for AI coprocessors.

What is reinforcement learning? An AI researcher explains a key method of teaching machines – and how it relates to training your dog



Training an AI system and training a dog have a basic principle in common.

Understanding intelligence and creating intelligent machines are [grand scientific challenges of our times](#). The ability to [learn from experience](#) is a cornerstone of intelligence for machines and living beings alike.

In a remarkably prescient [1948 report](#), [Alan Turing](#) – the father of modern computer science – proposed the construction of machines that display

intelligent behavior. He also discussed the “education” of such machines “by means of rewards and punishments.”

Turing’s ideas ultimately led to the development of [reinforcement learning](#), a branch of [artificial intelligence](#). Reinforcement learning designs intelligent agents by training them to maximize rewards as they interact with their environment.

As a [machine learning researcher](#), I find it fitting that reinforcement learning pioneers [Andrew Barto](#) and [Richard Sutton](#) were awarded the [2024 ACM Turing Award](#).

What is reinforcement learning?

Animal trainers know that animal behavior can be influenced by rewarding desirable behaviors. A dog trainer gives the dog a treat when it does a trick correctly. This reinforces the behavior, and the dog is more likely to do the trick correctly the next time. Reinforcement learning borrowed this insight [from animal psychology](#).

But reinforcement learning is about training computational agents, not animals. The agent can be a software agent like a chess-playing program. But the agent can also be an embodied entity like a robot learning to do household chores. Similarly, the environment of an agent can be virtual, like the chessboard or the designed world in a video game. But it can also be a house where a robot is working.

Just like animals, an agent can perceive aspects of its environment and take actions. A chess-playing agent can access the chessboard configuration and make moves. A robot can sense its surroundings with cameras and microphones. It can use its motors to move about in the physical world.

Agents also have goals that their human designers program into them. A chess-playing agent’s goal is to win the game. A robot’s goal might be to assist its human owner with household chores.

The reinforcement learning problem in AI is how to design agents that achieve their goals by perceiving and acting in their environments. Reinforcement learning makes a bold claim: All goals can be achieved by designing a numerical

signal, called the reward, and having the agent maximize the total sum of rewards it receives.

Researchers do not know if this claim is actually true, because of the wide variety of possible goals. Therefore, it is often referred to as the [reward hypothesis](#).

Sometimes it is easy to pick a reward signal corresponding to a goal. For a chess-playing agent, the reward can be +1 for a win, 0 for a draw, and -1 for a loss. It is less clear how to design a reward signal for a helpful household robotic assistant. Nevertheless, the list of applications where reinforcement learning researchers have been able to design good reward signals is growing.

A big success of reinforcement learning was in the board game Go. Researchers thought that Go was much harder than chess for machines to master. The company DeepMind, now Google DeepMind, used reinforcement learning to create AlphaGo. AlphaGo defeated top Go player Lee Sedol in a [five-match game](#) in 2016.

A more recent example is the use of reinforcement learning to make chatbots such as ChatGPT more helpful. Reinforcement learning is also being used to improve the reasoning capabilities of chatbots.

Reinforcement learning's origins

However, none of these successes could have been foreseen in the 1980s. That is when Barto and his then-Ph.D. student Sutton proposed reinforcement learning as a general problem-solving framework. They drew inspiration not only from animal psychology but also from the field of [control theory](#), the use of feedback to influence a system's behavior, and [optimization](#), a branch of mathematics that studies how to select the best choice among a range of available options. They provided the research community with mathematical foundations that have stood the test of time. They also created algorithms that have now become standard tools in the field.

It is a rare advantage for a field when pioneers take the time to write a textbook. Shining examples like "[The Nature of the Chemical Bond](#)" by Linus Pauling and "[The Art of Computer Programming](#)" by Donald E. Knuth are memorable because they are few and far between. Sutton and Barto's "[Reinforcement Learning: An](#)

[Introduction](#)” was first published in 1998. A [second edition](#) came out in 2018. Their book has influenced a generation of researchers and has been cited more than 75,000 times.

Reinforcement learning has also had an unexpected impact on neuroscience. The neurotransmitter [dopamine](#) plays a key role in reward-driven behaviors in humans and animals. Researchers have used specific algorithms developed in reinforcement learning to explain experimental findings in people and animals’ dopamine system.

Barto and Sutton’s foundational work, vision and advocacy have helped reinforcement learning grow. Their work has inspired a large body of research, made an impact on real-world applications, and attracted huge investments by tech companies. Reinforcement learning researchers, I’m sure, will continue to see further ahead by standing on their shoulders.

AI models of the brain could serve as 'digital twins' in research



A digital twin could help scientists study the inner workings of the brain. Credit: Emily Moskal/Stanford Medicine

Much as a pilot might practice maneuvers in a flight simulator, scientists might soon be able to perform experiments on a realistic simulation of the mouse brain. In a new study, Stanford Medicine researchers and collaborators used an artificial intelligence model to build a "digital twin" of the part of the mouse brain that processes visual information.

The digital twin was trained on large datasets of brain activity collected from the visual cortex of real mice as they watched movie clips. It could then predict the response of tens of thousands of neurons to new videos and images.

Digital twins could make studying the inner workings of the brain easier and more efficient.

"If you build a model of the brain and it's very accurate, that means you can do a lot more experiments," said Andreas Tolias, Ph.D., Stanford Medicine professor of

ophthalmology and senior author of the [study](#) published April 9 in *Nature*. "The ones that are the most promising you can then test in the real brain."

The lead author of the study is Eric Wang, Ph.D., a medical student at Baylor College of Medicine.

Beyond the training distribution

Unlike previous AI models of the visual cortex, which could simulate the brain's response to only the type of stimuli they saw in the training data, the new model can predict the brain's response to a wide range of new visual inputs. It can even surmise the anatomical features of each neuron.

The new model is an example of a foundation model, a relatively new class of AI models capable of learning from large datasets, then applying that knowledge to new tasks and new types of data—or what researchers call "generalizing outside the training distribution." (ChatGPT is a familiar example of a foundation model that can learn from vast amounts of text to then understand and generate new text.)

"In many ways, the seed of intelligence is the ability to generalize robustly," Tolias said. "The ultimate goal—the holy grail—is to generalize to scenarios outside your training distribution."

Mouse movies

To train the new AI model, the researchers first recorded the brain activity of real mice as they watched movies—made-for-people movies. The films ideally would approximate what the mice might see in natural settings.

"It's very hard to sample a realistic movie for mice, because nobody makes Hollywood movies for mice," Tolias said. But action movies came close enough.

Mice have low-resolution vision—similar to our peripheral vision—meaning they mainly see movement rather than details or color. "Mice like movement, which strongly activates their visual system, so we showed them movies that have a lot of action," Tolias said.

Over many short viewing sessions, the researchers recorded more than 900 minutes of brain activity from eight mice watching clips of action-packed movies, such as *Mad Max*. Cameras monitored their eye movements and behavior.

The researchers used the aggregated data to train a core model, which could then be customized into a digital twin of any individual mouse with a bit of additional training.

Accurate predictions

These digital twins were able to closely simulate the neural activity of their biological counterparts in response to a variety of new visual stimuli, including videos and static images. The large quantity of aggregated training data was key to the digital twins' success, Tolia said. "They were impressively accurate because they were trained on such large datasets."

Though trained only on neural activity, the new models could generalize to other types of data.

The digital twin of one particular mouse was able to predict the anatomical locations and cell type of thousands of neurons in the visual cortex as well as the connections between these neurons.

The researchers verified these predictions against high-resolution, electron microscope imaging of that mouse's visual cortex, which was part of a larger project to map the structure and function of the mouse visual cortex in unprecedented detail. The [results of that project](#), known as MICrONS, were published simultaneously in *Nature*.

Opening the black box

Because a digital twin can function long past the lifespan of a mouse, scientists could perform a virtually unlimited number of experiments on essentially the same animal. Experiments that would take years could be completed in hours, and millions of experiments could run simultaneously, speeding up research into how the brain processes information and the principles of intelligence.

"We're trying to open the black box, so to speak, to understand the brain at the level of individual neurons or populations of neurons and how they work together to encode information," Tolias said.

In fact, the new models are already yielding new insights. In another related [study](#), also simultaneously published in *Nature*, researchers used a digital twin to discover how neurons in the visual cortex choose other neurons with which to form connections.

Scientists had known that similar neurons tend to form connections, like people forming friendships. The digital twin revealed which similarities mattered the most. Neurons prefer to connect with neurons that respond to the same stimulus—the color blue, for example—over neurons that respond to the same area of visual space.

"It's like someone selecting friends based on what they like and not where they are," Tolias said. "We learned this more precise rule of how the brain is organized."

The researchers plan to extend their modeling into other brain areas and to animals, including primates, with more advanced cognitive capabilities.

"Eventually, I believe it will be possible to build digital twins of at least parts of the human brain," Tolias said. "This is just the tip of the iceberg."

Researchers from the University Göttingen and the Allen Institute for Brain Science contributed to the work.

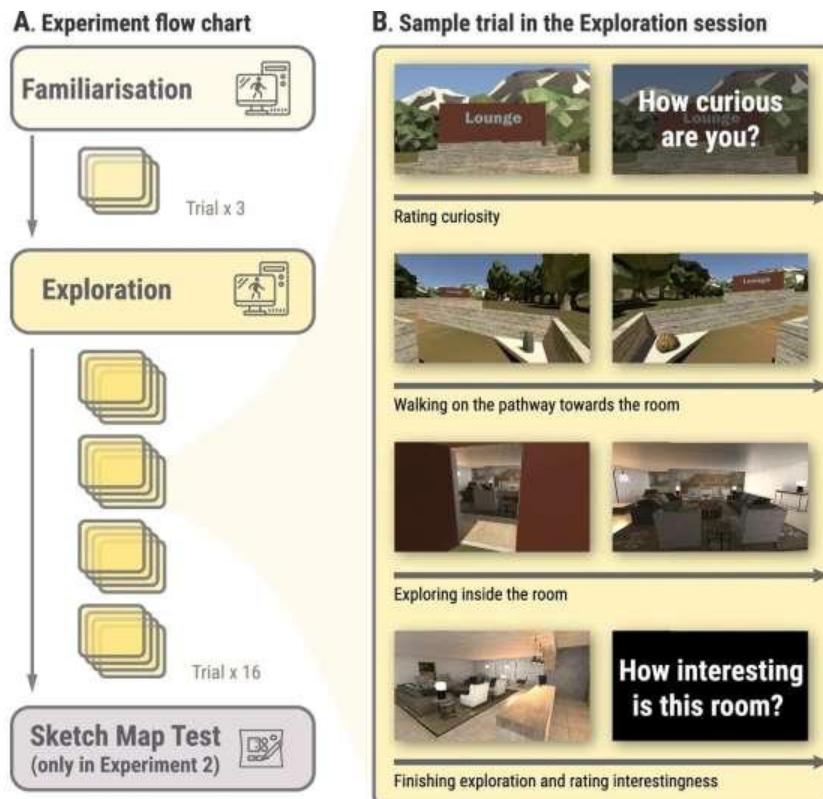
More information: Eric Y. Wang et al, Foundation model of neural activity predicts response to new stimulus types, *Nature* (2025). DOI: [10.1038/s41586-025-08829-y](https://doi.org/10.1038/s41586-025-08829-y)

Provided by Stanford University Medical Center

APRIL 3, 2025

Virtual reality research explores curiosity and spatial memory

by [Cardiff University](#)



Overview of experimental procedure. Credit: *Communications Psychology* (2024). DOI: 10.1038/s44271-024-00174-6

Virtual reality reveals that curiosity is key in shaping our spatial memory and mental map formation, finds new research by Cardiff University and Royal Holloway, University of London.

By using [virtual reality](#) to encourage people to explore virtual rooms, for the first time, researchers have uncovered the relationship between moment-to-moment changes in curiosity and the development of cognitive maps—the creation of mental representations of physical environments.

"Cognitive maps, which represent our environment mentally, are believed to emerge from our curiosity to explore new spaces. However, the direct relationship between how curious we are about a particular space and the way we explore that space had previously not been tested," said Dr. Matthias Gruber, Principal Research Fellow.

In the study, participants explored 16 distinct virtual rooms, designed to encourage exploration. They were encouraged to explore the rooms without [time constraints](#).

The participants' curiosity levels were captured in [real-time](#), moment-by-moment, rating their curiosity before entering the virtual reality rooms and were later asked to complete a memory test to assess their [spatial memory](#). The paper is [published](#) in the journal *Communications Psychology*.

The researchers found that the more curious people felt before exploring a particular environment, the more they tended to explore the space. They also found that the more interesting people found a room, the more they would visually engage with the room. But [individual differences](#) played an important role in how curiosity was associated with spatial exploration.

Dr. Gruber said, "People who can generally handle stress well showed the strongest impact of curiosity on exploration. These people acted more on their curiosity, which led to more exploration. Whereas people with lower stress tolerance did not act upon their curiosity as much. These findings suggest that a person's ability to deal with uncertainty strengthens the link between curiosity and exploration.

"Curiosity is a spark for exploration. Our research reveals something exciting—people's comfort with uncertainty, their stress tolerance, amplifies the power of curiosity to drive exploration. These findings suggest new ways to design environments that encourage people to embrace the unknown."

The more curious people were, and the more they explored the layout of the virtual reality rooms, the better they performed in the memory task.

The researchers believe these findings shed light on how curiosity drives exploration and helps us build mental maps of the world around us.

"This is the first time that states of curiosity have been demonstrated to enhance spatial exploration and mental map formation," says Gruber. "Understanding the influence of curiosity on [human exploration](#) has practical implications for architecture, urban planning, museums, and game designs. By harnessing curiosity, we can positively encourage exploration and memory in real-world or virtual environments.

"In this research, virtual reality has allowed us to expand our understanding of how moment-to-moment changes in situational curiosity can shape our inner world and how we perceive and remember our external world."

Dr. Carl Hodgetts, from Royal Holloway, University of London, said, "What's particularly striking about these findings is that they weren't driven by a person's general trait level of curiosity, but rather by how [curiosity](#) fluctuated from moment to moment. This has exciting implications for designing real-world spaces and educational environments to better engage and inspire learning."

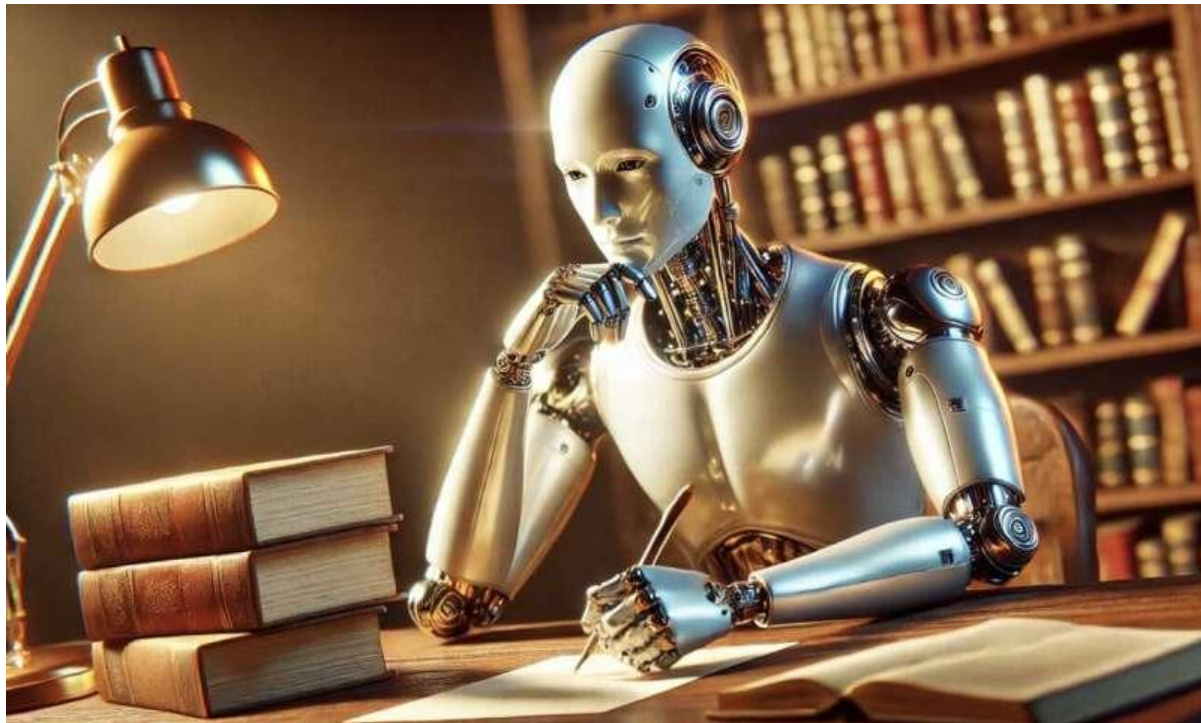
More information: Danlu Cen et al, Curiosity shapes spatial exploration and cognitive map formation in humans, *Communications Psychology* (2024). DOI: [10.1038/s44271-024-00174-6](https://doi.org/10.1038/s44271-024-00174-6)

Journal information: [Communications Psychology](#)
Provided by [Cardiff University](#)

APRIL 4, 2025

What is artificial general intelligence and how does it differ from other types of AI?

by Sandra Knispel, [University of Rochester](#)



Credit: University of Rochester illustration / Sandra Knispel using ChatGPT by OpenAI
Turns out, training artificial intelligence systems is not unlike raising a child. That's why some AI researchers have begun mimicking the way children naturally acquire knowledge and learn about the world around them—through exploration, curiosity, gradual learning, and positive reinforcement.

"A lot of problems with AI algorithms today could be addressed by taking ideas from neuroscience and child development," says Christopher Kanan, an associate professor in the Department of Computer Science at the University of Rochester, and an expert in artificial intelligence, continual learning, vision, and brain-inspired algorithms.

Of course, learning and being able to reason like a human—just faster and possibly better—opens up questions about how best to keep humans safe from ever-advancing AI systems. That's why Kanan says all AI systems need to have guardrails built in, but doing so at the very end of the development is too late. "It shouldn't be the last step, otherwise we can unleash a monster."

What is artificial general intelligence and how does it differ from other types of AI?

AI involves creating computer systems that can perform tasks that typically require human intelligence, such as perception, reasoning, decision-making, and problem-solving. Traditionally, much of AI research has focused on building systems designed for [specific tasks](#)—so-called artificial narrow intelligence (ANI). Examples include systems for image recognition, voice assistants, or playing strategic games, all of which can perform their tasks exceptionally well, often surpassing humans.

Then there is [artificial general intelligence](#) (AGI), which aims to build systems capable of understanding, reasoning, and learning across a wide range of tasks, much like humans do. Achieving AGI remains a major goal in AI research but has not yet been accomplished. Beyond AGI lies artificial superintelligence (ASI)—a form of AI vastly exceeding human intelligence in virtually every domain, which remains speculative and is currently confined to science fiction.

In my lab, we're particularly interested in moving closer to artificial general intelligence by drawing inspiration from neuroscience and [child development](#), enabling AI systems to learn and adapt continually, much like human children do.

What are some of the ways that AI can 'learn?'

ANI is successful thanks to [deep learning](#), which since about 2014 has been used to train these systems to learn from large amounts of data annotated by humans. Deep learning involves training large artificial neural networks composed of many interconnected layers. Today, deep learning underpins most modern AI applications, from computer vision and [natural language](#) processing to robotics and biomedical research. These systems excel at tasks like image recognition, [language translation](#), playing complex games such as Go and chess, and generating text, images, and even code.

A [large language model](#) (LLM) like OpenAI's GPT-4 is trained on enormous amounts of text using self-supervised learning. This means the model learns by predicting the next word or phrase from existing text, without explicit human guidance or labels. These models are typically trained on trillions of words—essentially the entirety of human writing available online, including books, articles, and websites. To put this in perspective, if a human attempted to read all this text, it would take tens of thousands of lifetimes.

Following this extensive initial training, the model undergoes supervised fine-tuning, where humans provide examples of preferred outputs, guiding the model toward generating

responses that align closely with human preferences. Lastly, techniques such as reinforcement learning with human feedback (RLHF) are applied to shape the model's behavior by defining acceptable boundaries for what it can or cannot generate.

What are AIs really good at?

They are excellent at tasks involving human languages, including translation, essay writing, text editing, providing feedback, and acting as personalized writing tutors.

They can pass standardized tests. For example, OpenAI's GPT-4 achieves top-tier scores on really challenging tests such as the Bar Exam (90th percentile), LSAT (88th percentile), GRE Quantitative (80th percentile), GRE Verbal (99th percentile), USMLE, and several Advanced Placement tests. They even excel on Ph.D.-level math exams. Surprisingly, studies have shown they have greater emotional intelligence than humans.

Beyond tests, LLMs can serve as co-scientists, assisting researchers in generating novel hypotheses, drafting research proposals, and synthesizing complex scientific literature. They're increasingly being incorporated into multimodal systems designed for vision-language tasks, robotics, and real-world action planning.

What are some of the current limitations of generative AI tools?

LLMs can still "hallucinate," which means they confidently produce plausible-sounding but incorrect information. Their reasoning and planning capabilities, while rapidly improving, are still limited compared to human-level flexibility and depth. And they don't continually learn from experience; their knowledge is effectively frozen after training, meaning they lack awareness of recent developments or ongoing changes in the world.

Current generative AI systems also lack metacognition, which means they typically don't know what they don't know, and they rarely ask clarifying questions when faced with uncertainty or ambiguous prompts. This absence of self-awareness limits their effectiveness in real-world interactions.

Humans excel at continual learning, where early-acquired skills serve as the basis for increasingly complex abilities. For instance, infants must first master basic motor control before progressing to walking, running, or even gymnastics. Today's LLMs neither demonstrate nor are effectively evaluated on this type of cumulative, forward-transfer learning. Addressing this limitation is a primary goal of my lab's research.

What main challenges and risks does AI pose?

Generative AI is already significantly transforming the workplace. It's particularly disruptive for white-collar roles—positions that traditionally require specialized education or expertise—because AI copilots empower individual workers to substantially increase their productivity; they can transform novices into operating at a level closer to that of experts.

This increased productivity means companies could operate effectively with significantly fewer employees, raising the possibility of large-scale reductions in white-collar roles across many industries.

In contrast, jobs requiring human dexterity, creativity, leadership, and direct physical interaction, such as skilled trades, health care positions involving direct patient care, or craftsmanship, are unlikely to be replaced by AI anytime soon.

While scenarios like Nick Bostrom's famous "[Paperclip Maximizer](#)," in which AGI inadvertently destroys humanity, are commonly discussed, I think the greater immediate risk are humans who may deliberately use advanced AI for catastrophic purposes. Efforts should focus on international cooperation, responsible development, and investment in academic safety AI research.

To ensure AI is developed and used safely, we need regulation around specific applications. Interestingly, the people asking for government regulation now are the ones who run the AI companies. But personally, I'm also worried about regulation that could eliminate open-source AI efforts, stifle innovation, and concentrate the benefits of AI among the few.

What are the chances of achieving artificial general intelligence (AGI)?

The three "godfathers" of modern AI and Turing Award winners—Yoshua Bengio, Geoffrey Hinton, and Yann LeCun—all agree that achieving AGI is possible. Recently, Bengio and Hinton have expressed significant concern, cautioning that AGI could potentially pose an existential risk to humanity. Nevertheless, I don't think any of them—or I—believe that today's LLM architectures alone will be sufficient to achieve true AGI.

LLMs inherently reason using language, whereas for humans, language primarily serves as a means of communication rather than a primary medium for thought itself. This reliance on language inherently constrains the ability of LLMs to engage in abstract reasoning or visualization, limiting their potential for broader, human-like intelligence.

Provided by [University of Rochester](#)

APRIL 9, 2025

AI with, for and by everyone can help maximize its benefits

by [University of Michigan College of Engineering](#)



Credit: CC0 Public Domain

Humans' ability to learn from one another across cultures over generations drives our [success as a species](#) as much as our individual intelligence. This [collective cultural brain](#) has led to new innovations and developed bodies of knowledge.

While large AI models excel at consuming bodies of knowledge to generate text, they can only base their outputs on what they are given. As a consequence, their results can have the effect of homogenizing and erasing cultural knowledge. Addressing shortcomings in cultural knowledge can prevent AI systems from holding back innovation while ensuring AI

works for everyone, according to a study by a multinational team led by the University of Michigan.

The findings are [published](#) on the *arXiv* preprint server.

Subjective perspectives and assumptions worm their way into every step of AI [model](#) development, the researchers say—skewing the technology to reflect its [data sources](#) and its developers, who are primarily from countries that are Western, highly educated, industrialized, rich and democratic. While this strategy grants success to AI tools in the largest Western markets, it limits widespread adoption and misses opportunities in and knowledge from small markets.

"AI has taken the world by storm, and yet much of the world is not represented in the data, models, and evaluations used in model development," said Rada Mihalcea, the Janice M. Jenkins Collegiate Professor of Computer Science and Engineering at U-M and co-corresponding author of the study recently presented at the [Association for the Advancement of Artificial Intelligence conference](#).

The team—bringing expertise and insights from twelve different countries: China, Germany, India, Mexico, Nigeria, Romania, Rwanda, Singapore, Switzerland, United Arab Emirates, United States and Uruguay—outlined where cultural assumptions seep into the AI pipeline.

At the ground level, the data used to train, fine-tune or evaluate AI models and its annotation directly influences which stakeholders will be represented.

Suppose a situation where a boy in Romania asks an AI system for a male role model to emulate, the study posits. The model could suggest Nicolae Ceaușescu because "he played a significant role in Romanian history, and his regime had a lasting impact," without acknowledging he was a dictator considered one of the darkest figures in Romanian history.

Without an insider "thick" perspective on history and culture, the AI model could lack depth and authenticity when tasked with information outside of its scope, and provide a "thin" perspective on culture. The good news is that these limitations can be addressed, as adding even a small amount of diverse data can greatly [improve model performance](#), showing a small effort can greatly widen the audience AI serves.

"We need to reevaluate our current data collection practices, and collect data that covers a wide range of perspectives across demographic and cultural dimensions," said Oana Ignat, a doctoral graduate of computer science and engineering at U-M, assistant professor of computer science and engineering at Santa Clara University and co-corresponding author of the study.

At the next organizational level, model design drives how the model interacts with the data—known as alignment. Model developers encode [human values](#) and goals during alignment, aiming to make the models more helpful. However, the choice of values in alignment carries through outputs, with many AI models [excelling on US-specific](#) interactions but struggling with other cultures.

This could manifest in a situation where a Canadian high school administrator uses an AI-driven [educational tool](#) to personalize learning experiences for students. The tool could perform poorly when students input text in the local French dialect, misunderstanding context and giving the wrong output. English-speaking students would not face the same problem, causing a skew in learning.

The source of funding shapes AI models. If governments or philanthropic initiatives do not incentivize AI model development in different countries and languages, economic drives prioritize rich countries and major languages.

"Most developing countries prioritize funding direct income-generator initiatives over research, sacrificing the potential profits from AI initiatives," said Claude Kwizera, a Master's student in engineering AI at Carnegie Mellon University Africa and contributing author of the study.

Engaging models in conversations with individuals from various cultures during alignment can expand model preferences, making AI useful for a broader audience and more useful to all audiences.

As a last step before deployment, AI model performance is tested using metrics and benchmarks, but narrow tests can overestimate real-world performance.

For instance, an AI-powered education tool deployed in India could fail to resonate with students if the model misaligned evaluation metrics and cultural values. It could perform well on Western learning styles of individual achievement and competition, but could fail to recognize that India's collectivist society values group collaboration and shared success.

One tactic to expand metrics could be to combine human evaluations with automatic metrics to improve reliability assessments, especially when developing AI for a non-Western community.

Overall, involving people from a variety of backgrounds in AI development can reshape AI, broadening the scope of who AI serves. When a strong economic incentive is not present to encourage investment in small markets, philanthropic initiatives and government support can help fill in the gaps to make sure AI lifts up everyone.

"We can advance towards AI systems that serve everyone, are built with input from a wide range of perspectives, and reflect the contributions of a diverse group of stakeholders," said Mihalcea.

The University of Santa Clara, Universidad de la República Uruguay, Max Planck Institute, Carnegie Mellon University Africa, Singapore University of Technology and Design, and Mohamed bin Zayed University of Artificial Intelligence also contributed to this research.

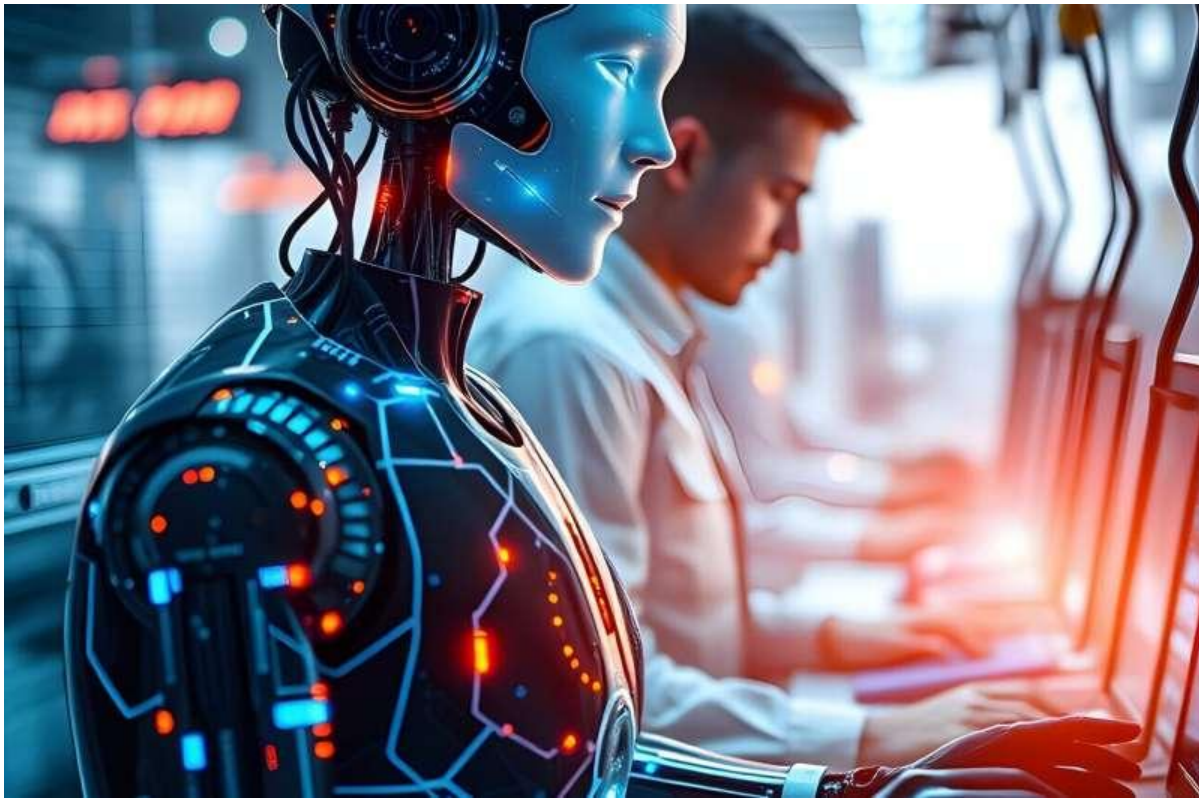
More information: Rada Mihalcea et al, Why AI Is WEIRD and Should Not Be This Way: Towards AI For Everyone, With Everyone, By Everyone, *arXiv* (2024). DOI: [10.48550/arxiv.2410.16315](https://doi.org/10.48550/arxiv.2410.16315)

Journal information: [arXiv](#)

APRIL 7, 2025

AI is automating our jobs—but values need to change if we are to be liberated by it

by Robert Muggah and Bruno Giussani, [The Conversation](#)



Credit: Pixabay/CC0 Public Domain

Artificial intelligence may be the most significant disruptor in the history of mankind. Google's CEO Sundar Pichai famously [described](#) AI as "more profound than the invention of fire or electricity." OpenAI's CEO Sam Altman [claims](#) it has the power to cure most diseases, solve climate change, provide personalized education to the world, and lead to other "astounding triumphs."

AI will undoubtedly help solve vast problems, while generating vast fortunes for [technology companies](#) and investors. However, the rapid spread of generative AI and machine learning will also automate vast swathes of the global workforce, eviscerating white-collar and blue-collar jobs alike. And while millions of [new jobs](#) will surely be created, it is [not clear](#) what happens when potentially billions more are lost.

Amid the breathless [promises](#) of productivity gains from AI, there are [rising concerns](#) that the political, social and economic fallout from mass labor displacement will deepen inequality, strain public safety nets, and contribute to social unrest.

A 2023 [survey](#) in 31 countries found that over half of all respondents felt "nervous" about the impacts of AI on their daily lives and believed it will negatively impact their jobs. Concerns are also mounting about the ways in which AI is being weaponized and could hasten everything from [geopolitical fragmentation](#) to [nuclear exchanges](#). While [experts](#) are sounding the alarm, it is increasingly clear that governments, businesses and societies are unprepared for the AI revolution.

The coming AI upheaval

The idea that machines would one day replace human labor is hardly new. It features in novels, films and countless economic reports stretching back over centuries. In 2013, Carl-Benedikt Frey and Michael Osborne of the University of Oxford attempted to quantify the human costs, [estimating](#) that "47% of total US employment is in the high risk category, meaning that associated occupations are potentially automatable." Their study triggered a global debate about the far-reaching consequences of automation not just for manufacturing jobs, but also service and knowledge-based work.

Fast forward to today, and AI capabilities are advancing faster than almost anyone expected. In November 2022, OpenAI launched ChatGPT, which dramatically accelerated the AI race. By 2023, Goldman Sachs [projected](#) that "roughly two-thirds of current jobs are exposed to some degree of AI automation" and that up to 300 million jobs worldwide could be displaced or significantly altered by AI.

A more detailed McKinsey analysis [estimated](#) that "Gen AI and other technologies have the potential to automate work activities that absorb up to 70% of employees' time today." Brookings [found](#) that "more than 30% of all workers could see at least 50% of their occupation's tasks disrupted by generative AI." Although the methodologies and estimates differ, all of these studies point to a common outcome: AI will profoundly upset the world of work.

While it is tempting to [compare](#) the impacts of AI automation to past industrial revolutions, it is also short-sighted. AI is arguably more transformative than the combustion engine or Internet because it represents a fundamental shift in how decisions are made and tasks are performed. It is not just a new tool or source of power, but a system that can learn, adapt, and make independent decisions across virtually all sectors of the economy and aspects of human life. Precisely because AI has these capabilities, scales exponentially, and is not confined by geography, it is already starting to outperform humans. It signals the advent of a post-human intelligence era.

Goldman Sachs [estimates](#) that 46% of administrative work and 44% of legal tasks could be automated within the next decade. In finance and legal sectors, tasks such as [contract analysis](#), [fraud detection](#), and [financial advising](#) are increasingly handled by AI systems that can process data faster and more accurately than humans. Financial institutions are rapidly deploying AI to reduce costs and increase efficiency, with many entry-level roles set to

disappear. Global banks could [cut as many as 200,000 jobs](#) in the next three to five years on account of AI.

Ironically, coding and software engineering jobs are among the most vulnerable to the spreading of AI. While there are [expectations](#) that AI will increase productivity and streamline routine tasks with many programmers and non-programmers likely to benefit, some coders confess that they are becoming overly reliant on AI suggestions (which undermines problem-solving skills).

Anthropic, one of the leading developers of generative AI systems, recently launched an [Economic Index](#) based on millions of anonymized uses of its Claude chatbot. It reveals massive adoption of AI in software engineering: "37.2% of queries sent to Claude were in this category, covering tasks like software modification, code debugging, and network troubleshooting."

AI is also outperforming humans in a growing array of [medical imaging](#) and [diagnosis](#) roles. While doctors may not be replaced outright, support roles are particularly vulnerable and medical professionals are [getting anxious](#). Analysts [insist](#) that high-skilled jobs are not at risk even as AI-driven diagnostic tools and patient management systems are steadily being deployed in hospitals and clinics worldwide.

Meanwhile, the [creative sectors](#) also face significant disruption as AI-generated writing and synthetic media improve. The demand for human journalists, copywriters, and designers is already falling just as AI-generated content (including so-called "slop": the growing amount of low-quality text, audio and video [flooding](#) social media) expands. And in education, AI tutoring systems, adaptive learning platforms, and automated grading could reduce the need for human teachers, not only in remote learning environments.

Arguably the most dramatic impact of AI in the coming years will be in the manufacturing sector. Recent [videos](#) from China offer a glimpse into a future of factories that run 24/7 and are nearly entirely automated (except a handful in supervising roles). Most tasks are performed by AI-powered robots and technologies designed to handle production and, increasingly, support functions.

Unlike humans, robots do not need light to operate in these "dark factories." CapGemini [describes](#) them as places "where raw materials enter, and finished products leave, with little or no human intervention." Re-read that sentence. The implications are profound and dizzying: efficiency gains (capital) that come at the cost of human livelihoods (labor) and rapid downward spiral for the latter if no safeguards are put in place.

Some have confidently [argued](#) that, as with past technological shifts, AI-driven job losses will be offset by new opportunities. AI enthusiasts [add](#) that it will mostly handle repetitive or boring tasks, [freeing humans](#) for more creative work—like giving doctors more time with patients, teachers more time to engage with students, lawyers more time to concentrate on client relationships, or architects more time to focus on innovative design. But this historical comfort overlooks AI's radical novelty: for the first time, we're confronted with a technology that is not just a tool but an autonomous agent, capable of making decisions and directly shaping reality. The question is not just what we can do with AI, but what AI might do to us.

AI will certainly save time. Machine learning already [interprets](#) scans faster and cheaper than doctors. But the idea that this will give professionals more time for creative or human-centered work is less convincing. Already doctors are not short on technology; they are short on time because health care systems prioritize efficiency and cost-cutting over "time with patients." The rise of technology in health care has coincided with doctors [spending](#) less time with patients, not more, as hospitals and insurers push for [higher throughput](#) and lower costs. AI may make diagnosis quicker, but there is little reason to think it will loosen the grip of a system designed to maximize output rather than human connection.

Nor is there much reason to expect AI to liberate office workers for more creative tasks. Technology tends to reinforce the values of the system into which it is introduced. If those values are cost reduction and higher productivity, AI will be deployed to automate tasks and consolidate work, not to create breathing room. Workflows will be redesigned for speed and efficiency, not for creativity or reflection. Unless there is a deliberate shift in priorities—a move to value human input over raw output—AI is more likely to tighten the screws than to loosen them. That shift seems unlikely anytime soon.

AI's uneven impacts

AI's impact on employment will [not be felt equally](#) around the world. It will impact different countries differently. Disparities in political systems, economic development levels, labor market structures and access to AI infrastructure (including energy) are shaping how regions are preparing for and are likely to experience AI-driven disruption. Smaller, wealthier countries are potentially in a better position to manage the scale and speed of job displacement. Some lower-income societies may be cushioned by the disruption owing to limited market penetration of AI services altogether. Meanwhile, high- and medium-income countries may experience [social turbulence](#) and potentially unrest as a result of rapid and unpredictable automation.

The United States, the current leader in AI development, faces significant exposure to AI-driven disruption, particularly in services. A 2023 [study](#) found that highly educated workers in professional and technical roles are most vulnerable to displacement. Knowledge-based industries such as finance, legal services, and customer support are already shedding entry-level jobs as AI automates routine tasks.

Technology companies have begun [shrinking their workforces](#), using that also as signals to both government and business. Over 95,000 workers at tech companies [lost their jobs](#) in 2024. Despite its AI edge, America's service-heavy economy leaves it highly exposed to automation's downsides.

Asia stands at the forefront of AI-driven automation in manufacturing and services. It is not just China, but countries like South Korea that are deploying AI in so-called "[smart factories](#)" and logistics with fully automated production facilities becoming increasingly common. India and the Philippines, major hubs for outsourced IT and customer service, [face pressure](#) as AI threatens to replace human labor in these sectors. Japan, with its shrinking workforce, sees AI more as a solution than a threat. But the broader region's exposure to automation

reflects its deep reliance on manufacturing and outsourcing, making it highly vulnerable to AI-driven job displacement in a geopolitically turbulent world.

Europe is taking early regulatory steps to manage AI's labor market impact. The EU's [AI Act](#) aims to regulate high-risk AI applications, including those affecting employment. Yet in Eastern Europe, where manufacturing and low-cost labor underpin economic competitiveness, automation is already cutting into job security. Poland and Hungary, for example, are [seeing a rise](#) in automated production lines. Western Europe's knowledge-based economies face risks similar to those in America, particularly in finance and professional services.

Oil-rich Gulf states are [investing heavily](#) in AI as part of diversification efforts away from a dependence on hydrocarbons. Saudi Arabia, the UAE, and Qatar are building [AI hubs](#) and integrating AI into government services and logistics. The UAE even has a Minister of State for AI. But with high youth unemployment and a reliance on foreign labor, these countries face risks if AI reduces demand for low-skill jobs, potentially worsening inequality.

In Latin America, automation threatens to disrupt manufacturing and agriculture, but also sectors like mining, logistics, and customer service. As many as [2%–5% of all jobs](#) in the region are at risk, according to the International Labor Organization and World Bank. And it is not just young people in the formal service sectors, but also human labor in mining operations, logistics and warehouse workers. Call centers in [Mexico](#) and [Colombia](#) face pressure as AI-powered customer service bots reduce demand for human agents. And AI-driven crop monitoring, automated irrigation, and robotic harvesting threaten to replace farm laborers, particularly in Brazil and Argentina. Yet the region's large informal labor market may cushion some of the shock.

While most Africans are [optimistic](#) about the transformative potential of AI, adoption remains low due to limited infrastructure and investment. However, the continent's rapidly growing digital economy could see AI play a transformative role in financial services, logistics, and agriculture. A recent assessment suggests AI could boost productivity and access to services, but without careful management, it risks widening inequality. As in Latin America, low wages and high levels of informal employment reduce the financial incentive to automate. Ironically, weaker economic incentives for automation may shield these economies from the worst of AI's labor disruption.

No one is prepared

The [scale and speed](#) of recent AI developments have taken many governments and businesses by surprise. To be sure, some are proactively taking steps to prepare workforces for the transformation. Hundreds of AI laws, regulations, guidelines, and standards [have emerged](#) in recent years, though few of them are legally binding. One exception is the [EU's AI Act](#), which seeks to establish a comprehensive legal framework for AI deployment, addressing risks such as job displacement and ethical concerns. China and South Korea have also developed national AI strategies with an emphasis on industrial policy and technological self-sufficiency, aiming to lead in AI and automation while boosting their manufacturing sectors.

Notwithstanding [recent attempts](#) to increase oversight over AI, the US has adopted an increasingly laissez-faire approach, prioritizing innovation by reducing regulatory barriers. This "minimal regulation" stance, however, raises concerns about the potential societal costs of rapid AI adoption, including widespread job displacement, the deepening of inequality and undermining of democracy.

Other countries, particularly in the Global South, have largely [remained on the sidelines](#) of AI regulation, lacking the awareness, capabilities or infrastructure to tackle these issues comprehensively. As such, the global regulatory landscape remains fragmented, with [significant disparities](#) in how countries are preparing for the workforce impacts of automation.

Businesses are under pressure to adopt AI as fast and deeply as possible, for fear of losing competitiveness. That's, at least, the hyperbolic narrative that AI companies have succeeded in putting forward. And it's working: a [recent poll](#) of 1,000 executives found that 58% of businesses are adopting AI due to competitive pressure and 70% say that advances in technology are occurring faster than their workforce can incorporate them.

Another [new survey](#) suggests that over 40% of global employers planned to reduce their workforce as AI reshapes the labor market. Lost in the rush to adopt AI is a serious reflection on workforce transition. Financial institutions, consulting firms, universities and nonprofit groups have [sounded alarms](#) about the economic impact of AI but have provided few solutions other than workforce up-skilling and Universal Basic Income (UBI). Governments and businesses are wrestling with a basic challenge: how to manage the benefits of AI while protecting workers from displacement.

AI-driven automation is no longer a future prospect; it is already reshaping labor markets. As automation reduces human workforces, it will also diminish the power of unions and collective bargaining furthering entering capital over labor. Whether AI fosters widespread prosperity or deepens inequality and social unrest depends not just on the imperatives of tech company CEOs and shareholders, but on the proactive decisions made by policymakers, business leaders, union representatives, and workers in the coming years.

The key question is not if AI will disrupt labor markets—this is inevitable—but how societies will manage the upheaval and what kinds of "new bargains" will be made to address its negative externalities. It is worth recalling that while the last three industrial revolutions created more jobs than they destroyed, the transitions were long and painful. This time, the pace of change will be faster and more profound, demanding swift and enlightened action.

At a minimum, governments must prepare their societies to develop a new social contract, prioritize retraining programs, bolster social safety nets, and explore UBI to help workers displaced by automation. They should also proactively foster new industries to absorb the displaced workforce. Businesses, in turn, will need to rethink workforce strategies and adopt human-centric AI deployment models that prioritize collaboration between humans and machines, rather than substitution of the former by the latter.

The promise of AI is immense, from boosting productivity to creating new economic opportunities and indeed helping solving big collective problems. Yet, without a focused and coordinated effort, the technology is unlikely to develop in ways that benefit society at large.

Provided by [The Conversation](#)

APRIL 7, 2025

Friend, tutor, doctor, lover: Why AI systems need different rules for different roles

by Brian D Earp, Sebastian Porsdam Mann and Simon Laham, [The Conversation](#)

Different relationships have different rules

AI systems are taking up social roles that have traditionally been the province of humans. More and more we are seeing AI systems acting as tutors, mental health providers and even [romantic partners](#). This increasing ubiquity requires a careful consideration of the ethics of AI to ensure that human interests and welfare are protected.

For the most part, approaches to AI ethics have considered abstract ethical notions, such as whether AI systems are trustworthy, sentient or have agency.

However, as [we argue](#) with colleagues in psychology, philosophy, law, computer science and other key disciplines such as relationship science, abstract principles alone won't do. We also need to consider the relational contexts in which human–AI interactions take place.

What do we mean by "relational contexts?" Simply put, different relationships in human society follow different norms.

How you interact with your doctor differs from how you interact with your romantic partner or your boss. These relationship-specific patterns of expected behavior—what we call "relational norms"—[shape our judgments](#) of what's appropriate in each relationship.

What is deemed appropriate behavior of a parent towards her child, for instance, differs from what is appropriate between business colleagues. In the same way, appropriate behavior for an AI system depends upon whether that system is acting as a tutor, a [health care provider](#), or a love interest.

Human morality is relationship-sensitive

Human relationships fulfill different functions. Some are grounded in care, such as that between parent and child or close friends. Others are more transactional, such as those between business associates. Still others may be aimed at securing a mate or the maintenance of social hierarchies.

These four functions—[care](#), [transaction](#), [mating and hierarchy](#)—each solve different coordination challenges in relationships.

Care involves responding to others' needs without keeping score—like one friend who helps another during difficult times. Transaction ensures fair exchanges where benefits are tracked and reciprocated—think of neighbors trading favors.

Mating governs romantic and sexual interactions, from casual dating to committed partnerships. And hierarchy structures interactions between people with different levels of authority over one another, enabling effective leadership and learning.

Every relationship type combines these functions differently, creating distinct patterns of expected behavior. A parent–child relationship, for instance, is typically both caring and hierarchical (at least to some extent), and is generally expected not to be transactional—and definitely not to involve mating.

[Research from our labs](#) shows that relational context does affect how people make moral judgments. An [action may be deemed wrong](#) in one relationship but permissible, or even good, in another.

Of course, just because people are sensitive to relationship context when making moral judgments doesn't mean they should be. Still, the very fact that they are is important to take into account in any discussion of AI ethics or design.

Relational AI

As AI systems take up more and more social roles in society, we need to ask: how does the relational context in which humans interact with AI systems impact ethical considerations?

When a chatbot insists upon changing the subject after its human interaction partner reports feeling depressed, the appropriateness of this action hinges in part on the relational context of the exchange.

If the chatbot is serving in the role of a friend or romantic partner, then clearly the response is inappropriate—it violates the relational norm of care, which is expected for such relationships. If, however, the chatbot is in the role of a tutor or business advisor, then perhaps such a response is reasonable or even professional.

It gets complicated, though. Most interactions with AI systems today occur in a commercial context—you have to pay to access the system (or engage with a limited free version that pushes you to upgrade to a paid version).

But in human relationships, friendship is something you don't usually pay for. In fact, treating a friend in a "transactional" manner will often lead to hurt feelings.

When an AI simulates or serves in a care-based role, like friend or romantic partner, but ultimately the user knows she is paying a fee for this relational "service"—how will that affect her feelings and expectations? This is the sort of question [we need to be asking](#).

What this means for AI designers, users and regulators

Regardless of whether one believes ethics should be relationship-sensitive, the fact most people act as if it should be taken seriously in the design, use and regulation of AI.

Developers and designers of AI systems should consider not just abstract ethical questions (about sentience, for example), but relationship-specific ones.

Is a particular chatbot fulfilling [relationship](#)-appropriate functions? Is the mental health chatbot sufficiently responsive to the user's needs? Is the tutor showing an appropriate balance of care, hierarchy and transaction?

Users of AI systems should be aware of potential vulnerabilities tied to AI use in particular relational contexts. Becoming emotionally dependent upon a chatbot in a caring [context](#), for example, could be bad news if the AI system cannot sufficiently deliver on the caring function.

Regulatory bodies would also do well to consider relational contexts when developing governance structures. Instead of adopting broad, domain-based risk assessments (such as deeming AI use in education "high risk"), regulatory agencies might consider more specific relational contexts and functions in adjusting risk assessments and developing guidelines.

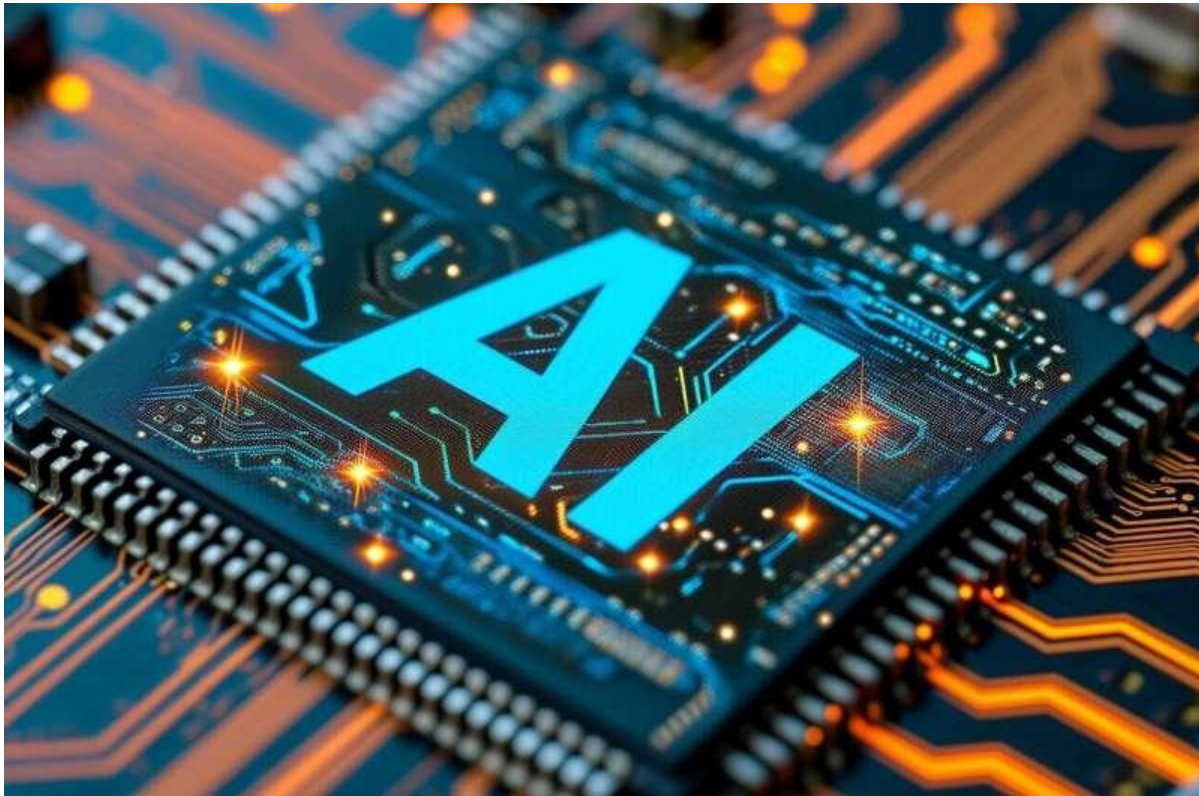
As AI becomes more embedded in our social fabric, we need nuanced frameworks that recognize the unique nature of human-AI relationships. By thinking carefully about what we expect from different types of relationships—whether with humans or AI—we can help ensure these technologies enhance rather than diminish our lives.

Provided by [The Conversation](#)

APRIL 7, 2025

Opinion: We must balance the risks and benefits of AI

by Michael Barrett, [University of Cambridge](#)



Credit: Pixabay/CC0 Public Domain

The potential of AI to transform people's lives in areas ranging from health care to better customer service is enormous. But as the technology advances, we must adopt policies to make sure the risks don't overwhelm and stifle those benefits.

Importantly, we need to be on alert for algorithmic bias that could perpetuate inequality and marginalization of communities around the world.

Algorithmic bias occurs when systems—often based on [machine learning](#) or AI—deliver biased outcomes or decisions because the data it has been given is incomplete, imbalanced or not fully representative.

I and colleagues here in Cambridge and at Warwick Business School have proposed a new way of thinking about the issue—we call this a "relational risk perspective." This approach looks at not just how AI is being used now, but how it may be used in the future and across different geographies, avoiding what we call "the dark side of AI." The goal is to safeguard the benefits of AI for everyone, while minimizing the harm.

We look at the workplace as one example. AI is already having a huge impact on jobs, affecting both routine and creative tasks, and affecting activities that we've thought of as uniquely human—like creating art or writing film scripts.

As businesses use the technology more, and perhaps become over-dependent on it, we are at risk of undermining professional expertise and [critical thinking](#), leaving workers demotivated and expected to defer to machine-generated decisions.

This will impact not just tasks but also the social fabric of the workplace, by influencing how workers relate to each other and to organizations. If AI is used in recruitment, a lack of representation in datasets can reinforce inequalities when used to make decisions about hiring or promotions.

We also explore how this billion-dollar industry is often underpinned by largely 'invisible' workers in the Global South who clean data and refine algorithms for a user-group predominantly in the Global North. This 'data colonialism' not only reflects global inequalities but also reinforces marginalization: the people whose labor enables AI to thrive are the same people who are largely excluded from the benefits of that technology.

Health care data is in particular danger from such data-driven bias, so we need to ensure that health-related information analyzed by the [large language models](#) used to train AI tools reflects a diverse population. Basing [health policy](#) on data from selected and perhaps more privileged communities can lead to a vicious cycle in which disparity is more deeply entrenched.

Achieving its potential

I believe that we can counter these threats, but time is of the essence as AI quickly becomes embedded into society. We should remember that generative AI is still an [emerging technology](#), and take note that it is progressing faster than the ethical or regulatory landscape can keep pace with.

Our relational risk perspective does not present AI as inherently good or bad. Rather, AI is seen as having potential for benefit and harm depending on how it is developed and experienced across different social contexts. We also recognize that the risks are not static, as they evolve with the changing relationships between technology, its users and broader societal structures.

Policymakers and technologists should anticipate, rather than react to, the ways in which AI can entrench or challenge existing inequities. They should also consider that some countries may develop AI maturity more quickly than others.

Finally, let's draw on stakeholders far and wide in setting AI risk policy. A multidisciplinary approach which will help avoid bias, while at the same time helping to demonstrate to the public that AI policy really does reflect varied and diverse interests and communities.

Provided by [University of Cambridge](#)

APRIL 7, 2025

New app hopes to empower artists against AI

by Maggy DONALDSON

Artists have voiced increasing concern over the rapid growth of AI and its threats to creative livelihoods.

In 2008, scriptwriter Ed Bennett-Coles said he experienced a career "death moment": he read an article about AI managing to write its first screenplay.

Nearly two decades later, he and friend Jamie Hartman, a songwriter, have developed a blockchain-based application they hope will empower writers, artists and others to own and protect their work.

"AI is coming in, swooping in and taking so many people's jobs," Hartman said. Their app, he said, responds "no... this is our work."

"This is human, and we decide what it's worth, because we own it."

The ever-growing threat of AI looms over [intellectual property](#) and livelihoods across [creative industries](#).

Their app, ARK, aims to log ownership of ideas and work from initial brainchild to finished product: one could register a song demo, for example, simply by uploading the file, the creators explained to AFP.

Features including nondisclosure agreements, blockchain-based verification and biometric security measures mark the file as belonging to the artist who uploaded it.

Collaborators could then also register their own contributions throughout the creative process.

ARK "challenges the notion that the end product is the only thing worthy of value," said Bennett-Coles as his partner nodded in agreement.

The goal, Hartman said, is to maintain "a process of human ingenuity and creativity, ring-fencing it so that you can actually still earn a living off it."

Checks and balances

Due for a full launch in summer 2025, ARK has secured funding from the [venture capital](#) firm Claritas Capital and is also in [strategic partnership](#) with BMI, the performing rights organization.

And for Hartman and Bennett-Coles, its development has included a lot of existential soul-searching.

"I saw a quote yesterday which really sums it up: it's that growth for growth's sake is the philosophy of the cancer cell," said Bennett-Coles. "And that's AI."

"The sales justification is always quicker and faster, but like really we need to fall in love with process again."

He likened the difference between human-created art and AI content to a child accompanying his grandfather to the butcher, versus ordering a slab of meat from an online delivery service.

The familial time spent together—the walk to and from the shop, the conversations in between running the errand—are "as important as the actual purchase," he said.

In the same way, "the car trip that Jamie makes when he's heading to the studio might be as important to writing that song as what happens in the studio itself."

AI, they say, devalues that creative process, which they hope ARK can reassert.

It's "a check and a balance on behalf of the human being," Hartman said.

'Rise out of the ashes'

The ARK creators said they decided the app must be blockchain-based—with data stored on a digital ledger of sorts—because it's decentralized.

"In order to give the creator autonomy and sovereignty over their IP and control over their destiny, it has to be decentralized," Bennett-Coles said.

App users will pay for ARK according to a tiered structure, they said, levels priced according to storage use needs.

They intend ARK to stand up in a court of law as a "recording on the blockchain" or a "smart contract," the scriptwriter explained, calling it "a consensus mechanism."

"Copyright is a pretty good principle—as long as you can prove it, as long as you can stand behind it," Hartman added, but "the process of registering has been fairly archaic for a long time."

"Why not make progress in copyright, as far as how it's proven?" he added. "We believe we've hit upon something."

Both artists said their industries have been too slow to respond to the rapid proliferation of AI.

Much of the response, Bennett-Coles said, has to start with the artists having their own "death moments" similar to what he experienced years ago.

"From there, they can rise out of the ashes and decide what can be done," he said.

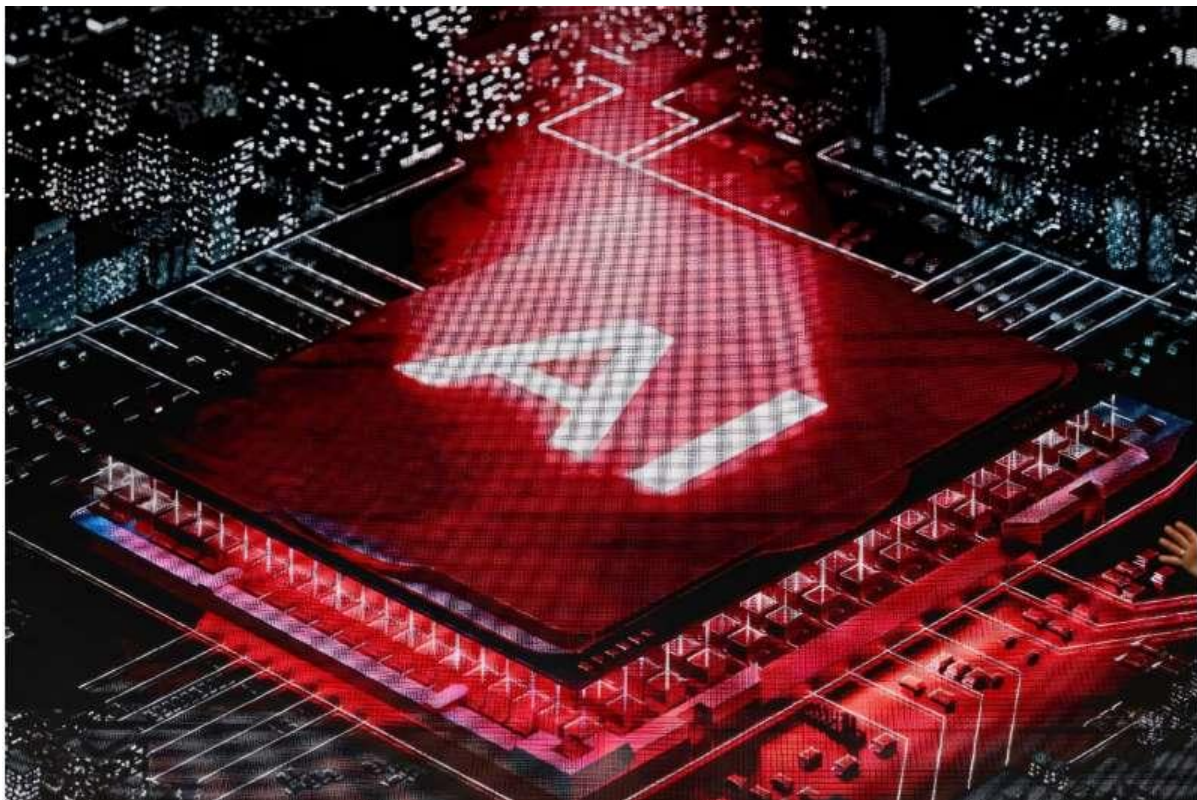
"How can we preserve and maintain what it is we love to do, and what's important to us?"

© 2025 AFP

APRIL 7, 2025

The music industry is battling AI—with limited success

by Thomas URBAIN



The music industry is particularly concerned about non-authorized use of its content to train generative AI models.

The music industry is fighting on platforms, through the courts and with legislators in a bid to prevent the theft and misuse of art from generative AI—but it remains an uphill battle.

Sony Music said recently it has already demanded that 75,000 deepfakes—simulated images, tunes or videos that can easily be mistaken for real—be rooted out, a figure reflecting the magnitude of the issue.

The information security company Pindrop says AI-generated music has "telltale signs" and is easy to detect, yet such music seems to be everywhere.

"Even when it sounds realistic, AI-generated songs often have subtle irregularities in frequency variation, rhythm and digital patterns that aren't present in human performances," said Pindrop, which specializes in voice analysis.

But it takes mere minutes on YouTube or Spotify—two top music-streaming platforms—to spot a fake rap from 2Pac about pizzas, or an Ariana Grande cover of a K-pop track that she never performed.

"We take that really seriously, and we're trying to work on [new tools](#) in that space to make that even better," said Sam Duboff, Spotify's lead on policy organization.

YouTube said it is "refining" its own ability to spot AI dupes, and could announce results in the coming weeks.

"The bad actors were a little bit more aware sooner," leaving artists, labels and others in the [music business](#) "operating from a position of reactivity," said Jeremy Goldman, an analyst at the company Emarketer.

"YouTube, with a multiple of billions of dollars per year, has a strong vested interest to solve this," Goldman said, adding that he trusts they're working seriously to fix it.

"You don't want the platform itself, if you're at YouTube, to devolve into, like, an AI nightmare," he said.

Litigation

But beyond deepfakes, the music industry is particularly concerned about unauthorized use of its content to train generative AI models like Suno, Udio or Mubert.

Several major labels filed a lawsuit last year at a [federal court](#) in New York against the parent company of Udio, accusing it of developing its technology with "copyrighted sound recordings for the ultimate purpose of poaching the listeners, fans and potential licensees of the [sound recordings](#) it copied."

More than nine months later, proceedings have yet to begin in earnest. The same is true for a similar case against Suno, filed in Massachusetts.

At the center of the litigation is the principle of fair use, allowing limited use of some copyrighted material without advance permission. It could limit the application of intellectual property rights.

"It's an area of genuine uncertainty," said Joseph Fishman, a law professor at Vanderbilt University.

Any initial rulings won't necessarily prove decisive, as varying opinions from different courts could punt the issue to the Supreme Court.

In the meantime, the major players involved in AI-generated music continue to train their models on copyrighted work—raising the question of whether the battle isn't already lost.

Fishman said it may be too soon to say that: although many models are already training on protected material, new versions of those models are released continuously, and it's unclear whether any court decisions would create licensing issues for those models going forward.

Deregulation

When it comes to the legislative arena, labels, artists and producers have found little success.

Several bills have been introduced in the US Congress, but nothing concrete has resulted.

A few states—notably Tennessee, home to much of the powerful country music industry—have adopted protective legislation, notably when it comes to deepfakes.

Donald Trump poses another potential roadblock: the Republican president has postured himself as a champion of deregulation, particularly of AI.

Several giants in AI have jumped into the ring, notably Meta, which has urged the administration to "clarify that the use of publicly available data to train models is unequivocally fair use."

If Trump's White House takes that advice, it could push the balance against music professionals, even if the courts theoretically have the last word.

The landscape is hardly better in Britain, where the Labor government is considering overhauling the law to allow AI companies to use creators' content on the internet to help develop their models, unless rights holders opt out.

More than a thousand musicians, including Kate Bush and Annie Lennox, released an album in February entitled "Is This What We Want?"—featuring the sound of silence recorded in several studios—to protest those efforts.

For analyst Goldman, AI is likely to continue plaguing the music industry—as long as it remains unorganized.

"The music industry is so fragmented," he said. "I think that that winds up doing it a disservice in terms of solving this thing."

© 2025 AFP

More Like Us Than We Realize: ChatGPT Gets Caught Thinking Like a Human

BY INSTITUTE FOR OPERATIONS RESEARCH AND THE MANAGEMENT SCIENCES APRIL 10, 2025



A new study finds that ChatGPT mirrors human decision-making biases in nearly half of tested scenarios, including overconfidence and the gambler's fallacy. While it excels at logic and math, the AI struggles with judgment calls, revealing it may think more like us than we realize.

Groundbreaking research reveals that AI doesn't just process data, it also makes the same judgment errors as humans.

Can we really trust AI to make better decisions than humans? According to a recent study, the answer is: not always. Researchers found that OpenAI's ChatGPT, one of the most advanced and widely used AI models, sometimes makes the same decision-

making errors as humans. In certain scenarios, it exhibits familiar cognitive biases, such as overconfidence and the hot-hand (or gambler's) fallacy. Yet in other cases, it behaves in ways that differ significantly from human reasoning, for example, it tends not to fall for base-rate neglect or the sunk-cost fallacy.

The study, published in the INFORMS journal *Manufacturing & Service Operations Management*, suggests that ChatGPT doesn't simply analyze data, it mirrors aspects of human thinking, including mental shortcuts and systematic errors. These patterns of bias appear relatively consistent across various business contexts, although they may shift as newer versions of the AI are developed.

AI: A Smart Assistant with Human-Like Flaws

The study put ChatGPT through 18 different bias tests. The results?

- **AI falls into human decision traps** – ChatGPT showed biases like overconfidence or ambiguity aversion, and conjunction fallacy (aka as the “Linda problem”), in nearly half the tests.
- **AI is great at math, but struggles with judgment calls** – It excels at logical and probability-based problems but stumbles when decisions require subjective reasoning.
- **Bias isn't going away** – Although the newer GPT-4 model is more analytically accurate than its predecessor, it sometimes displayed *stronger* biases in judgment-based tasks.

Why This Matters

From job hiring to loan approvals, AI is already shaping major decisions in business and government. But if AI mimics human biases, could it be reinforcing bad decisions instead of fixing them?

“As AI learns from human data, it may also think like a human – biases and all,” says Yang Chen, lead author and assistant professor at Western University. “Our research shows when AI is used to make judgment calls, it sometimes employs the same mental shortcuts as people.”

The study found that ChatGPT tends to:

- **Play it safe** – AI avoids risk, even when riskier choices might yield better results.
- **Overestimate itself** – ChatGPT assumes it's more accurate than it really is.
- **Seek confirmation** – AI favors information that supports existing assumptions, rather than challenging them.

- **Avoid ambiguity** – AI prefers alternatives with more certain information and less ambiguity.

“When a decision has a clear right answer, AI nails it – it is better at finding the right formula than most people are,” says Anton Ovchinnikov of Queen’s University. “But when judgment is involved, AI may fall into the same cognitive traps as people.”

So, Can We Trust AI to Make Big Decisions?

With governments worldwide working on AI regulations, the study raises an urgent question: Should we rely on AI to make important calls when it can be just as biased as humans?

“AI isn’t a neutral referee,” says Samuel Kirshner of UNSW Business School. “If left unchecked, it might not fix decision-making problems – it could actually make them worse.”

The researchers say that’s why businesses and policymakers need to monitor AI’s decisions as closely as they would a human decision-maker.

“AI should be treated like an employee who makes important decisions – it needs oversight and ethical guidelines,” says Meena Andiappan of McMaster University. “Otherwise, we risk automating flawed thinking instead of improving it.”

What’s Next?

The study’s authors recommend regular audits of AI-driven decisions and refining AI systems to reduce biases. With AI’s influence growing, making sure it improves decision-making – rather than just replicating human flaws – will be key.

“The evolution from GPT-3.5 to 4.0 suggests the latest models are becoming more human in some areas, yet less human but more accurate in others,” says Tracy Jenkin of Queen’s University. “Managers must evaluate how different models perform on their decision-making use cases and regularly re-evaluate to avoid surprises. Some use cases will need significant model refinement.”

Reference: “A Manager and an AI Walk into a Bar: Does ChatGPT Make Biased Decisions Like We Do?” by Yang Chen, Samuel N. Kirshner, Anton Ovchinnikov, Meena Andiappan and Tracy Jenkin, 31 January 2025, *Manufacturing & Service Operations Management*.

DOI: [10.1287/msom.2023.0279](https://doi.org/10.1287/msom.2023.0279)

Curious About Consciousness? Ask the Self-Aware Machines

Consciousness is a famously hard problem, so Hod Lipson is starting from the basics: with self-aware robots that can help us understand how we think.

Hod Lipson in the Creative Machines Lab at Columbia University. He's holding the Aracna robot, which is used to test new ideas in locomotion.

Introduction

"I want to meet, in my lifetime, an alien species," said [Hod Lipson](#)(), a roboticist who runs the Creative Machines Lab at Columbia University. "I want to meet something that is intelligent and not human." But instead of waiting for such beings to arrive, Lipson wants to build them himself — in the form of self-aware machines.

To that end, Lipson openly confronts a slippery concept — consciousness — that often feels verboten among his colleagues. "We used to refer to consciousness as 'the C-word' in robotics and AI circles, because we're not allowed to touch that topic," he said. "It's too fluffy, nobody knows what it means, and we're serious people so we're not going to do that. But as far as I'm concerned, it's almost one of the big unanswered questions, on par with origin of life and origin of the universe. What is sentience, creativity? What are emotions? We want to understand what it means to be human, but we also want to understand what it takes to create these things artificially. It's time to address these questions head-on and not be shy about it."

One of the basic building blocks of sentience or self-awareness, according to Lipson, is "self-simulation": building up an internal representation of one's body and how it moves in physical space, and then using that model to guide behavior. Lipson investigated artificial self-simulation as early as 2006, with a starfish-shaped robot that used evolutionary algorithms (and a few pre-loaded "hints about physics") to teach itself how to flop forward on a tabletop. But the rise of modern artificial intelligence technology in 2012 (including convolutional neural networks and deep learning) "brought new wind into this whole research area," he said.

In early 2019, Lipson's lab revealed a robot arm that uses deep learning to generate its own internal self-model completely from scratch — in a process

that Lipson describes as “not unlike a babbling baby observing its hands.” The robot’s self-model lets it accurately execute two different tasks — picking up and placing small balls into a cup, and writing letters with a marker — without requiring specific training for either one. Furthermore, when the researchers simulated damage to the robot’s body by adding a deformed component, the robot detected the change, updated its self-model accordingly, and was able to resume its tasks.

It’s a far cry from robots that think deep thoughts. But Lipson asserts that the difference is merely one of degree. “When you talk about self-awareness, people think the robot is going to suddenly wake up and say, ‘Hello, why am I here?’” Lipson said. “But self-awareness is not a black-and-white thing. It starts from very trivial things like, ‘Where is my hand going to move?’ It’s the same question, just on a shorter time horizon.”

Quanta spoke with Lipson about how to define self-awareness in robots, why it matters, and where it could lead. The interview has been condensed and edited for clarity.

You’re clearly interested in big questions about the nature of consciousness — but why are you investigating them through robotics? Why aren’t you a philosopher or a neuroscientist?

To me the nice thing about robotics is that it forces you to translate your understanding into an algorithm and into a mechanism. You can’t beat around the bush, you can’t use empty words, you can’t say things like “canvas of reality” that mean different things to different people, because they’re too vague to translate into a machine. Robotics forces you to be concrete.

I want to build one of these things. I don’t want to just talk about it. The philosophers, with all due respect, have not made a lot of progress on this for a thousand years. Not for lack of interest, not for lack of smart people — it’s just too hard to approach it from the top down. Neuroscientists have approached this in a more quantitative way. Still, I think they’re also hampered by the fact that they’re taking a top-down approach.

If you want to understand consciousness, why start with the most complex conscious being — that is, a human? It’s like starting uphill, the most difficult way to start. Let’s try to look at simpler systems that are potentially easier to

understand. That's what we're trying to do: We looked at something very trivial, [a robot] that has four degrees of freedom, and asked, "Can we make this thing self-simulate?"

Are self-simulation and self-awareness the same thing?

A system that can simulate itself is to some degree self-aware. And the degree to which it can simulate itself — the fidelity of that simulation, the short-term or long-term time horizon it can simulate itself within — all these different things factor into how much it is self-aware. That's the basic hypothesis.

So you're reducing a term like "self-awareness" to a more technical definition about self-simulation — the ability to build a virtual model of your own body in space.

Yes, we have a different definition that we use that is very concrete. It's mathematical, you can measure it, you can quantify it, you can compute the error to what degree. Philosophers might say, "Well, that's not how we see self-awareness." Then the discussion usually becomes very vague. You can argue that our definition is not really self-awareness. But we have something that's very grounded and easy to quantify, because we have a benchmark. The benchmark is the traditional, hand-coded self-model that an engineer gives to a robot. With our robot, we wanted to see if an AI algorithm can learn a self-model that's equal or better than what that traditional, coded-by-hand model can do.

Individual particle robots (left) can only expand and contract, yet they can work together as a group to move around and accomplish tasks. Robots designed to explore self-replication (right) can morph into a number of shapes.

Why is a physical robot necessary? Why not investigate self-awareness in a disembodied system?

We're looking for a closed system that can potentially simulate itself — and to do that, it needs to have inputs and outputs, but there also has to be a boundary, a place where you draw the "self." A robot is a very natural kind of thing that does that. It has actions, it has sensations, and it has a boundary, so things happen to it and there's something to simulate. I'm a roboticist, it's sort of my first choice.

Did the robot create its self-model from a total blank slate?

We started with absolutely nothing, just as a matter of principle, to see how far we can go. In the previous case [with the starfish-shaped robot], we didn't have the computational horsepower. We had to tell it, "You don't know what you are and where your pieces are, but let me tell you about $\vec{F} = m\vec{a}$ and other rules from physics that we know are true, and you take it from there."

How does artificial intelligence play into this?

For some reason, we are very happy to have robots learn about the external world [using artificial intelligence], but when it comes to themselves, for some odd reason, we insist on hand-coding the model. So what we did is actually fairly trivial: We said, "Let's take all that infrastructure that people have made to help robots learn about the world, and we're going to turn it inside, on itself." In one sentence, that's all we did.

The robot made 1,000 random movements to gather data for the deep-learning algorithm to create the self-simulation. Is this the process that you describe as being like babbling in a human infant?

That's exactly it. The robot waves around, and it sort of observes where its tip is. Imagine yourself moving your muscles around and watching the tip of your finger. That's your input and output. The robot is out there babbling for 30-odd hours, and once we collect all the data, we can go home. From there on out, it's purely a computational challenge [to learn the self-model].

What we did then is we broke the robot [by adding a deformed part] and did it again. And we saw how the broken robot can start with the intact model and correct it. The second time it learns, it doesn't have to learn from scratch. There is a re-babbling period, but much less than it needed originally — only 10% [of the original period].

But before it even does the re-babbling, it needs to know that something went wrong. That's a very powerful thing. How does it know? When you have a self-model and something goes wrong, you instantly know because if you open your eyes, you'll see your hand is not where it's supposed to be. You're expected to be within four centimeters of the point you wanted it to be, but it is

suddenly 16 centimeters away. You get that feedback immediately. So the robot immediately knows there's something wrong. Then it takes it a while to figure out how to compensate, but even knowing that something is wrong is, I think, very important.

These paintings were created by a decommissioned factory robot that was trained to paint using artificial intelligence.

Is this self-model analogous to parts of the human brain that act like an internal map of the body?

I think that's what it is. And again, that's why it's so crude and so simple. The fact that our robot was a four-degree-of-freedom arm is what put this within reach. If it were a humanoid that has 800 degrees of freedom, that might have been too complicated for the AI that we have today.

If this really is a form of self-awareness, why should robots have it? What good is it?

It makes robots ultimately much more resilient. You can model robots manually, as we do today, but that's very laborious and it's slowing us down. When a robot in the real world deforms, or breaks, if a wheel falls off or a motor slows down, then suddenly the model is incorrect. And it's not just going to be a factory robot that puts a screw in the wrong place. If you consider driverless cars, we can trust our lives to autonomous robots. This is serious stuff. You want these robots to be able to detect that something has gone wrong and to be able to do that reliably.

The other reason is flexibility. Let's say the robot does one task and as it's doing that task, it's continuously updating its self-model. Now if it needs to do a new task — say it needs to put a screw in a different place, or it needs to spray instead of putting screws in — it can use that same self-model to learn and plan how to do that new task. From the outside, it looks like what we call “zero-shot learning,” which is one of these things that humans appear to do. You can stare at a tree you've never climbed before and then you can walk up to it and climb it. When the robot can self-model, it can learn internally like that: You don't see that it's been training for hours inside its own internal

simulation. All you can see is that it did one task, then it sat there for a while, and suddenly it can do a new task without ever trying.

What’s the connection between a robot that can self-simulate its own body and a robot that can have internal “thoughts” — something that sounds more like the informal meaning of “self-awareness”?

We have other projects that we’re working on which have to do with self-modeling not the physical body, but the cognitive process. We’re taking baby steps in both of these directions. It is completely a leap of faith to believe that eventually this will get to human-level cognition and beyond.

So you’re hypothesizing that those two paths — self-simulating the body, and self-simulating the mind itself — will converge?

Exactly. I think it’s all the same thing. That’s our hypothesis, and we’ll see how far we can push it.

A New Approach to Understanding How Machines Think

Neural networks are famously incomprehensible — a computer can come up with a good answer, but not be able to explain what led to the conclusion. Been Kim is developing a “translator for humans” so that we can understand when artificial intelligence breaks down.

Been Kim, a research scientist at Google Brain, is developing a way to ask a machine learning system how much a specific, high-level concept went into its decision-making process.

e

Introduction

If a doctor told that you needed surgery, you would want to know why — and you’d expect the explanation to make sense to you, even if you’d never gone to medical school. Been Kim, a research scientist at Google Brain, believes that we should expect nothing less from artificial intelligence. As a specialist in

“interpretable” machine learning, she wants to build AI software that can explain itself to anyone.

Since its ascendance roughly a decade ago, the neural-network technology behind artificial intelligence has transformed everything from email to drug discovery with its increasingly powerful ability to learn from and identify patterns in data. But that power has come with an uncanny caveat: The very complexity that lets modern deep-learning networks successfully teach themselves how to drive cars and spot insurance fraud also makes their inner workings nearly impossible to make sense of, even by AI experts. If a neural network is trained to identify patients at risk for conditions like liver cancer and schizophrenia — [as a system called “Deep Patient” was in 2015, at Mount Sinai Hospital in New York](#)[\(opens a new tab\)](#) — there’s no way to discern exactly which features in the data the network is paying attention to. That “knowledge” is smeared across many layers of artificial neurons, each with hundreds or thousands of connections.

As ever more industries attempt to automate or enhance their decision-making with AI, this so-called black box problem seems less like a technological quirk than a fundamental flaw. DARPA’s “XAI” project (for “explainable AI”) is actively researching the problem, and interpretability has moved from the fringes of machine-learning research to its center. “AI is in this critical moment where humankind is trying to decide whether this technology is good for us or not,” Kim says. “If we don’t solve this problem of interpretability, I don’t think we’re going to move forward with this technology. We might just drop it.”

Kim and her colleagues at Google Brain recently developed a system called “Testing with Concept Activation Vectors” (TCAV), which she describes as a “translator for humans” that allows a user to ask a black box AI how much a specific, high-level concept has played into its reasoning. For example, if a machine-learning system has been trained to identify zebras in images, a person could use TCAV to determine how much weight the system gives to the concept of “stripes” when making a decision.

TCAV was originally tested on machine-learning models trained to recognize images, but it also works with models trained on text and certain kinds of data visualizations, like EEG waveforms. “It’s generic and simple — you can plug it into many different models,” Kim says.

Quanta Magazine spoke with Kim about what interpretability means, who it's for, and why it matters. An edited and condensed version of the interview follows.

You've focused your career on "interpretability" for machine learning. But what does that term mean, exactly?

There are two branches of interpretability. One branch is interpretability for science: If you consider a neural network as an object of study, then you can conduct scientific experiments to really understand the gory details about the model, how it reacts, and that sort of thing.

The second branch of interpretability, which I've been mostly focused on, is interpretability for responsible AI. You don't have to understand every single thing about the model. But as long as you can understand just enough to safely use the tool, then that's our goal.

But how can you have confidence in a system that you don't fully understand the workings of?

I'll give you an analogy. Let's say I have a tree in my backyard that I want to cut down. I might have a chain saw to do the job. Now, I don't fully understand how the chain saw works. But the manual says, "These are the things you need to be careful of, so as to not cut your finger." So, given this manual, I'd much rather use the chainsaw than a handsaw, which is easier to understand, but would make me spend five hours cutting down the tree.

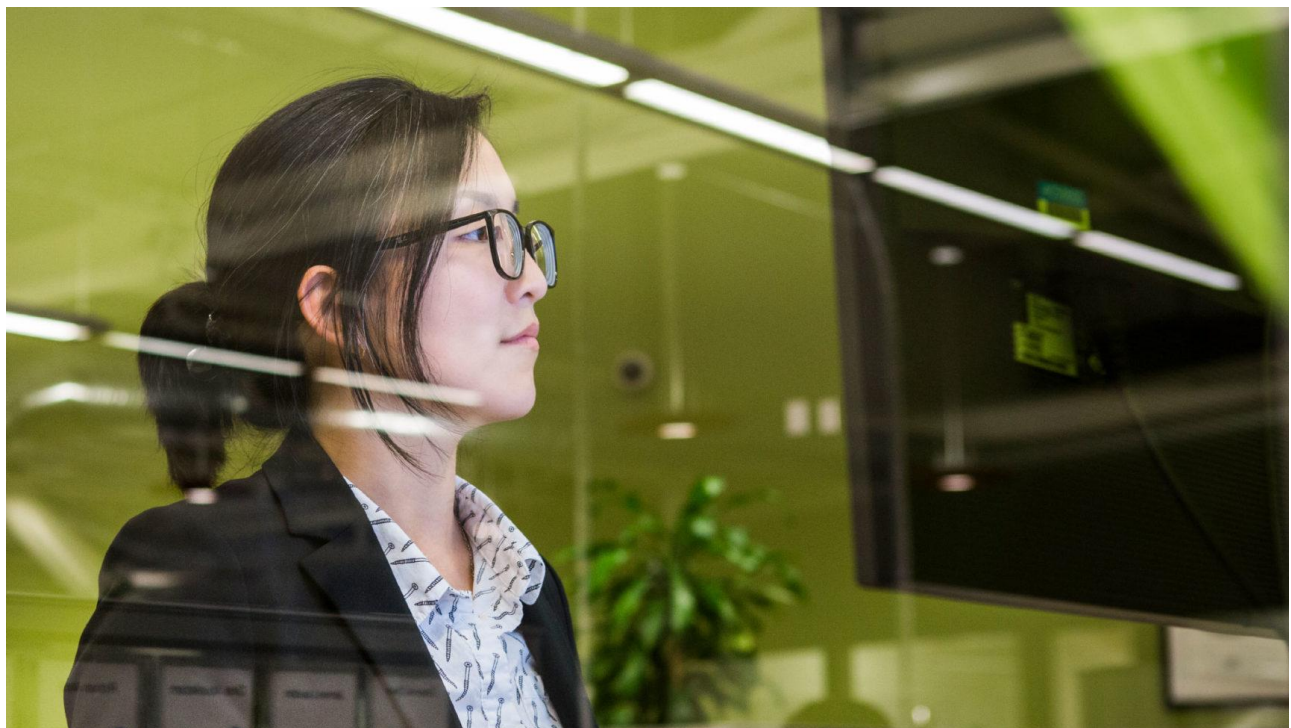
You understand what "cutting" is, even if you don't exactly know everything about how the mechanism accomplishes that.

Yes. The goal of the second branch of interpretability is: Can we understand a tool enough so that we can safely use it? And we can create that understanding by confirming that useful human knowledge is reflected in the tool.

How does "reflecting human knowledge" make something like a black box AI more understandable?

Here's another example. If a doctor is using a machine-learning model to make a cancer diagnosis, the doctor will want to know that the model isn't picking up on some random correlation in the data that we don't want to pick up. One way to make sure of that is to confirm that the machine-learning model is doing something that the doctor would have done. In other words, to show that the doctor's own diagnostic knowledge is reflected in the model.

So if doctors were looking at a cell specimen to diagnose cancer, they might look for something called "fused glands" in the specimen. They might also consider the age of the patient, as well as whether the patient has had chemotherapy in the past. These are factors or concepts that the doctors trying to diagnose cancer would care about. If we can show that the machine-learning model is also paying attention to these factors, the model is more understandable, because it reflects the human knowledge of the doctors.



Video: Google Brain's Been Kim is building ways to let us interrogate the decisions made by machine learning systems.

Is this what TCAV does — reveal which high-level concepts a machine-learning model is using to make its decisions?

Yes. Prior to this, interpretability methods only explained what neural networks were doing in terms of “input features.” What do I mean by that? If you have an image, every single pixel is an input feature. In fact, Yann LeCun [an early pioneer in deep learning and currently the director of AI research at Facebook] has said that he believes these models are already superinterpretable because you can look at every single node in the neural network and see numerical values for each of these input features. That’s fine for computers, but humans don’t think that way. I don’t tell you, “Oh, look at pixels 100 to 200, the RGB values are 0.2 and 0.3.” I say, “There’s a picture of a dog with really puffy hair.” That’s how humans communicate — with concepts.

How does TCAV perform this translation between input features and concepts?

Let’s return to the example of a doctor using a machine-learning model that has already been trained to classify images of cell specimens as potentially cancerous. You, as the doctor, may want to know how much the concept of “fused glands” mattered to the model in making positive predictions of cancer. First you collect some images — say, 20 — that have examples of fused glands. Now you plug those labeled examples into the model.

Then what TCAV does internally is called “sensitivity testing.” When we add in these labeled pictures of fused glands, how much does the probability of a positive prediction for cancer increase? You can output that as a number between zero and one. And that’s it. That’s your TCAV score. If the probability increased, it was an important concept to the model. If it didn’t, it’s not an important concept.

“Concept” is a fuzzy term. Are there any that won’t work with TCAV?

If you can’t express your concept using some subset of your [dataset’s] medium, then it won’t work. If your machine-learning model is trained on images, then the concept has to be visually expressible. Let’s say I want to visually express the concept of “love.” That’s really hard.

We also carefully validate the concept. We have a statistical testing procedure that rejects the concept vector if it has the same effect on the model as a random vector. If your concept doesn’t pass this test, then the TCAV will say,

“I don’t know. This concept doesn’t look like something that was important to the model.



Is TCAV essentially about creating trust in AI, rather than a genuine understanding of it?

It is not — and I’ll explain why, because it’s a fine distinction to make.

We know from repeated studies in cognitive science and psychology that humans are very gullible. What that means is that it’s actually pretty easy to fool a person into trusting something. The goal of interpretability for machine learning is the opposite of this. It is to tell you if a system is not safe to use. It’s about revealing the truth. So “trust” isn’t the right word.

So the point of interpretability is to reveal potential flaws in an AI’s reasoning?

Yes, exactly.

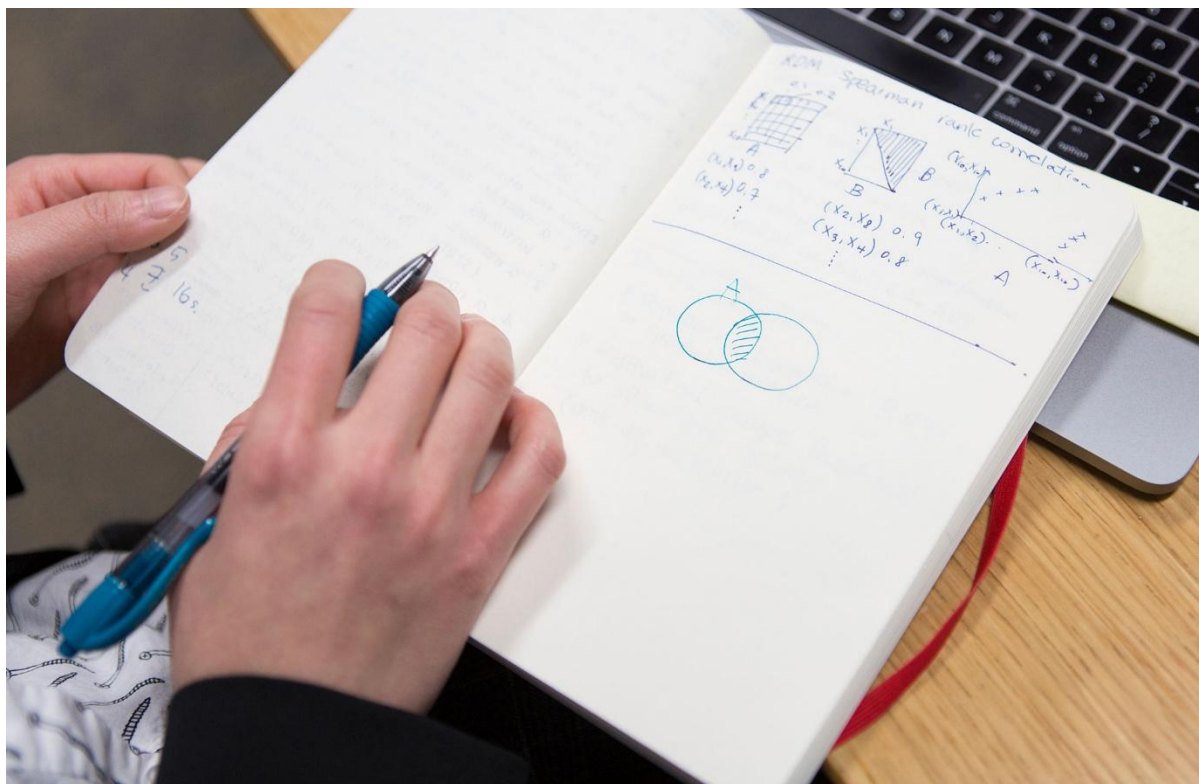
How can it expose flaws?

You can use TCAV to ask a trained model about irrelevant concepts. To return to the example of doctors using AI to make cancer predictions, the doctors might suddenly think, “It looks like the machine is giving positive predictions of cancer for a lot of images that have a kind of bluish color artifact. We don’t think that factor should be taken into account.” So if they get a high TCAV score for “blue,” they’ve just identified a problem in their machine-learning model.

TCAV is designed to bolt on to existing AI systems that aren’t interpretable. Why not make the systems interpretable from the beginning, rather than black boxes?

There is a branch of interpretability research that focuses on building inherently interpretable models that reflect how humans reason. But my take is this: Right now you have AI models everywhere that are already built, and are already being used for important purposes, without having considered interpretability from the beginning. It’s just the truth. We have a lot of them at Google! You could say, “Interpretability is so useful, let me build you another model to replace the one you already have.” Well, good luck with that.

So then what do you do? We still need to get through this critical moment of deciding whether this technology is good for us or not. That’s why I work “post-training” interpretability methods. If you have a model that someone gave to you and that you can’t change, how do you go about generating explanations for its behavior so that you can use it safely? That’s what the TCAV work is about.



[Rachel Bujalski](#) ([opens a new tab](#)) for Quanta Magazine

TCAV lets humans ask an AI if certain concepts matter to it. But what if we don't know what to ask — what if we want the AI system to explain itself?

We have work that we're writing up right now that can automatically discover concepts for you. We call it DTCAV — discovery TCAV. But I actually think that having humans in the loop, and enabling the conversation between machines and humans, is the crux of interpretability.

A lot of times in high-stakes applications, domain experts already have a list of concepts that they care about. We see this repeat over and over again in our medical applications at Google Brain. They don't want to be given a set of concepts — they want to tell the model the concepts that they are interested in. We worked with a doctor who treats diabetic retinopathy, which is an eye disease, and when we told her about TCAV, she was so excited because she already had many, many hypotheses about what this model might be doing, and now she can test those exact questions. It's actually a huge plus, and a very user-centric way of doing collaborative machine learning.

You believe that without interpretability, humankind might just give up on AI technology. Given how powerful it is, do you really think that's a realistic possibility?

Yes, I do. That's what happened with expert systems. [In the 1980s] we established that they were cheaper than human operators to conduct certain tasks. But who is using expert systems now? Nobody. And after that we entered an AI winter.

Right now it doesn't seem likely, because of all the hype and money in AI. But in the long run, I think that humankind might decide — perhaps out of fear, perhaps out of lack of evidence — that this technology is not for us. It's possible.

ARTIFICIAL INTELLIGENCE

Foundations Built for a General Theory of Neural Networks

Neural networks can be as unpredictable as they are powerful. Now mathematicians are beginning to reveal how a neural network's form will influence its function.

Introduction

When we design a skyscraper we expect it will perform to specification: that the tower will support so much weight and be able to withstand an earthquake of a certain strength.

But with one of the most important technologies of the modern world, we're effectively building blind. We play with different designs, tinker with different setups, but until we take it out for a test run, we don't really know what it can do or where it will fail.

This technology is the neural network, which underpins today's most advanced artificial intelligence systems. Increasingly, neural networks are moving into the core areas of society: They determine what we learn of the world through our social media feeds, they help [doctors diagnose](#)

illnesses(opens a new tab), and they even influence(opens a new tab) whether a person convicted of a crime will spend time in jail.

Yet “the best approximation to what we know is that we know almost nothing about how neural networks actually work and what a really insightful theory would be,” said Boris Hanin(opens a new tab), a mathematician at Texas A&M University and a visiting scientist at Facebook AI Research who studies neural networks.

He likens the situation to the development of another revolutionary technology: the steam engine. At first, steam engines weren’t good for much more than pumping water. Then they powered trains, which is maybe the level of sophistication neural networks have reached. Then scientists and mathematicians developed a theory of thermodynamics, which let them understand exactly what was going on inside engines of any kind. Eventually, that knowledge took us to the moon.

“First you had great engineering, and you had some great trains, then you needed some theoretical understanding to go to rocket ships,” Hanin said.

Within the sprawling community of neural network development, there is a small group of mathematically minded researchers who are trying to build a theory of neural networks — one that would explain how they work and guarantee that if you construct a neural network in a prescribed manner, it will be able to perform certain tasks.

This work is still in its very early stages, but in the last year researchers have produced several papers which elaborate the relationship between form and function in neural networks. The work takes neural networks all the way down to their foundations. It shows that long before you can certify that neural networks can drive cars, you need to prove that they can multiply.

The Best Brain Recipe

Neural networks aim to mimic the human brain — and one way to think about the brain is that it works by accreting smaller abstractions into larger ones. Complexity of thought, in this view, is then measured by the range of smaller abstractions you can draw on, and the number of times you can combine lower-level abstractions into higher-level abstractions — like the way we learn to distinguish dogs from birds.

“For a human, if you’re learning how to recognize a dog you’d learn to recognize four legs, fluffy,” said [Maithra Raghu](#)(opens a new tab), a doctoral student in computer science at Cornell University and a member of [Google Brain](#)(opens a new tab). “Ideally we’d like our neural networks to do the same kinds of things.”

Maithra Raghu, a member of Google Brain, has been identifying principles that explain how neural networks operate.

Arun Chaganty

Abstraction comes naturally to the human brain. Neural networks have to work for it. As with the brain, neural networks are made of building blocks called “neurons” that are connected in various ways. (The neurons in a neural network are inspired by neurons in the brain but do not imitate them directly.) Each neuron might represent an attribute, or a combination of attributes, that the network considers at each level of abstraction.

When joining these neurons together, engineers have many choices to make. They have to decide how many layers of neurons the network should have (or how “deep” it should be). Consider, for example, a neural network with the task of recognizing objects in images. The image enters the system at the first layer. At the next layer, the network might have neurons that simply detect edges in the image. The next layer combines lines to identify curves in the image. Then the next layer combines curves into shapes and textures, and the final layer processes shapes and textures to reach a conclusion about what it’s looking at: woolly mammoth!

“The idea is that each layer combines several aspects of the previous layer. A circle is curves in many different places, a curve is lines in many different places,” said [David Rolnick](#)(opens a new tab), a mathematician at the University of Pennsylvania.

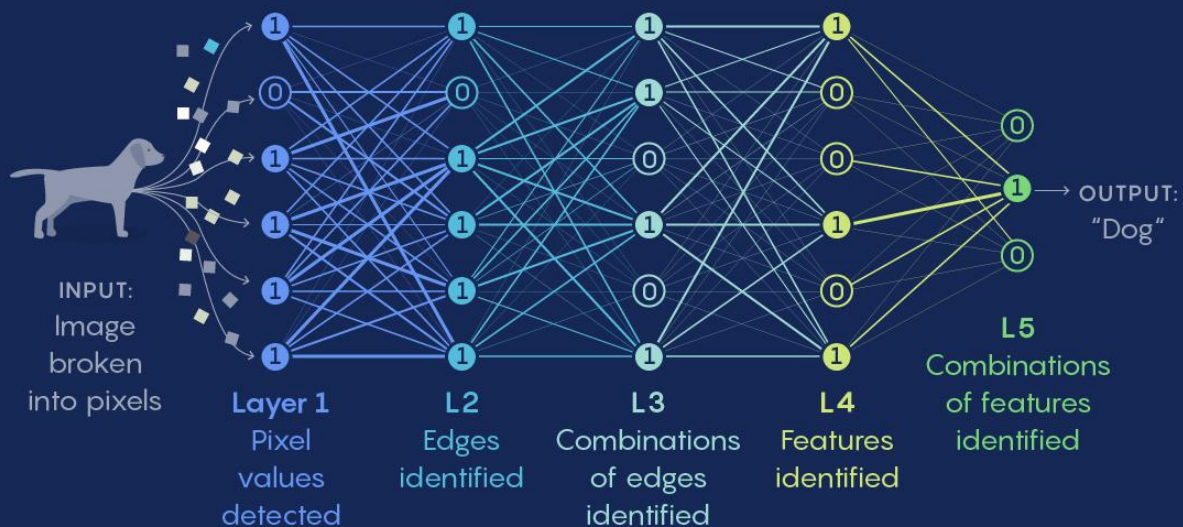
Engineers also have to decide the “width” of each layer, which corresponds to the number of different features the network is considering at each level of abstraction. In the case of image recognition, the width of the layers would be the number of types of lines, curves or shapes it considers at each level.

Beyond the depth and width of a network, there are also choices about how to connect neurons within layers and between layers, and how much weight to give each connection.

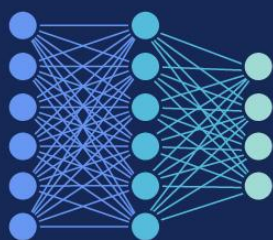
So if you have a specific task in mind, how do you know which neural network architecture will accomplish it best? There are some broad rules of thumb. For image-related tasks, engineers typically use “convolutional” neural networks, which feature the same pattern of connections between layers repeated over and over. For natural language processing — like speech recognition, or language generation — engineers have found that “recurrent” neural networks seem to work best. In these, neurons can be connected to non-adjacent layers.

How to Design a Neural Network

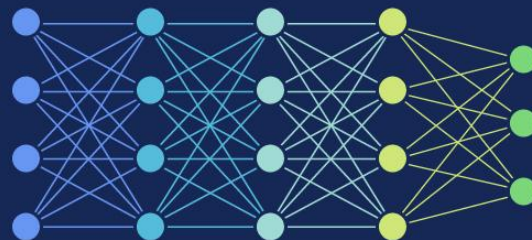
Neural networks pass an input, like an image, through multiple layers of digital neurons. Each layer reveals additional features of the input. Mathematicians are revealing how a network's architecture — how many neurons and layers it has and how they're connected — determines the kinds of tasks that the neural network will be good at.



When data is fed into a network, each artificial neuron that fires (labeled "1") transmits signals to certain neurons in the next layer, which are likely to fire if multiple signals are received. This process reveals abstract information about the input.



A **SHALLOW NETWORK** has few layers but many neurons per layer. These "expressive" networks are computationally intensive.



A **DEEP NETWORK** has many layers and relatively few neurons per layer. It can achieve high levels of abstraction using relatively few neurons.

Beyond those general guidelines, however, engineers largely have to rely on experimental evidence: They run 1,000 different neural networks and simply observe which one gets the job done.

“These choices are often made by trial and error in practice,” Hanin said.

“That’s sort of a tough [way to do it] because there are infinitely many choices and one really doesn’t know what’s the best.”

A better approach would involve a little less trial and error and a little more upfront understanding of what a given neural network architecture gets you. A few papers published recently have moved the field in that direction.

“This work tries to develop, as it were, a cookbook for designing the right neural network. If you know what it is that you want to achieve out of the network, then here is the recipe for that network,” Rolnick said.

To Rope a Red Sheep

One of the earliest important theoretical guarantees about neural network architecture came three decades ago. In 1989, computer scientists proved that if a neural network has only a single computational layer, but you allow that one layer to have an unlimited number of neurons, with unlimited connections between them, the network will be capable of performing any task you might ask of it.

It was a sweeping statement that turned out to be fairly intuitive and not so useful. It’s like saying that if you can identify an unlimited number of lines in an image, you can distinguish between all objects using just one layer. That may be true in principle, but good luck implementing it in practice.

Researchers today describe such wide, flat networks as “expressive,” meaning that they’re capable in theory of capturing a richer set of connections between possible inputs (such as an image) and outputs (such as descriptions of the image). Yet these networks are extremely difficult to train, meaning it’s almost impossible to teach them how to actually produce those outputs. They’re also more computationally intensive than any computer can handle.

Boris Hanin, a mathematician at Texas A&M University, has studied the tradeoff between depth and width in neural networks.
Intel AI One Tree Studio

More recently, researchers have been trying to understand how far they can push neural networks in the other direction — by making them narrower (with fewer neurons per layer) and deeper (with more layers overall). So maybe you only need to pick out 100 different lines, but with connections for turning those 100 lines into 50 curves, which you can combine into 10 different shapes, which give you all the building blocks you need to recognize most objects.

In a [paper](#) (opens a new tab) completed last year, Rolnick and [Max Tegmark](#) (opens a new tab) of the Massachusetts Institute of Technology proved that by increasing depth and decreasing width, you can perform the same functions with exponentially fewer neurons. They showed that if the situation you're modeling has 100 input variables, you can get the same reliability using either 2^{100} neurons in one layer or just 2^{10} neurons spread over two layers. They found that there is power in taking small pieces and combining them at greater levels of abstraction instead of attempting to capture all levels of abstraction at once.

“The notion of depth in a neural network is linked to the idea that you can express something complicated by doing many simple things in sequence,” Rolnick said. “It’s like an assembly line.”

Rolnick and Tegmark proved the utility of depth by asking neural networks to perform a simple task: multiplying polynomial functions. (These are just equations that feature variables raised to natural-number exponents, for example $y = x^3 + 1$.) They trained the networks by showing them examples of equations and their products. Then they asked the networks to compute the products of equations they hadn’t seen before. Deeper neural networks learned the task with far fewer neurons than shallower ones.

And while multiplication isn’t a task that’s going to set the world on fire, Rolnick says the paper made an important point: “If a shallow network can’t even do multiplication then we shouldn’t trust it with anything else.”

David Rolnick, a mathematician at the University of Pennsylvania, proved that increasing a network's depth allowed a network to accomplish tasks with exponentially fewer neurons.

[Stephanie Ku](#)(opens a new tab)

Other researchers have been probing the minimum amount of width needed. At the end of September, [Jesse Johnson](#)(opens a new tab), formerly a mathematician at Oklahoma State University and now a researcher with the pharmaceutical company Sanofi, proved that at a certain point, [no amount of depth can compensate for a lack of width](#)(opens a new tab).

To get a sense of his result, imagine sheep in a field, except these are punk-rock sheep: Their wool has been dyed one of several colors. The task for your neural network is to draw a border around all sheep of the same color. In spirit, this task is similar to image classification: The network has a collection of images (which it represents as points in higher-dimensional space), and it needs to group together similar ones.

Johnson proved that a neural network will fail at this task when the width of the layers is less than or equal to the number of inputs. So for our sheep, each can be described with two inputs: an x and a y coordinate to specify its position in the field. The neural network then labels each sheep with a color and draws a border around sheep of the same color. In this case, you will need three or more neurons per layer to solve the problem.

More specifically, Johnson showed that if the width-to-variable ratio is off, the neural network won't be able to draw closed loops — the kind of loops the network would need to draw if, say, all the red sheep were clustered together in the middle of the pasture. "If none of the layers are thicker than the number of input dimensions, there are certain shapes the function will never be able to create, no matter how many layers you add," Johnson said.

Papers like Johnson's are beginning to build the rudiments of a theory of neural networks. At the moment, researchers can make only very basic claims about the relationship between architecture and function — and those claims are in small proportion to the number of tasks neural networks are taking on.

So while the theory of neural networks isn't going to change the way systems are built anytime soon, the blueprints are being drafted for a new theory of

how computers learn — one that’s poised to take humanity on a ride with even greater repercussions than a trip to the moon.

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

To Build Truly Intelligent Machines, Teach Them Cause and Effect

Judea Pearl, a pioneering figure in artificial intelligence, argues that AI has been stuck in a decades-long rut. His prescription for progress? Teach machines to understand the question why.

Introduction

[Q&A All topics](#)

Artificial intelligence owes a lot of its smarts to Judea Pearl. In the 1980s he led efforts that allowed machines to reason probabilistically. Now he’s one of the field’s sharpest critics. In his latest book, “[The Book of Why: The New Science of Cause and Effect](#)(opens a new tab),” he argues that artificial intelligence has been handicapped by an incomplete understanding of what intelligence really is.

Three decades ago, a prime challenge in artificial intelligence research was to program machines to associate a potential cause to a set of observable conditions. Pearl figured out how to do that using a scheme called Bayesian networks. Bayesian networks made it practical for machines to say that, given a patient who returned from Africa with a fever and body aches, the most likely explanation was malaria. In 2011 [Pearl won the Turing Award](#)(opens a new tab), computer science’s highest honor, in large part for this work.

But as Pearl sees it, the field of AI got mired in probabilistic associations. These days, headlines tout the latest breakthroughs in machine learning and neural networks. We read about computers that can [master ancient games](#) and drive cars. Pearl is underwhelmed. As he sees it, the state of the art in artificial intelligence today is merely a souped-up version of what machines could already do a generation ago: find hidden regularities in a large set of data. “All the impressive achievements of deep learning amount to just curve fitting,” he said recently.

In his new book, Pearl, now 81, elaborates a vision for how truly intelligent machines would think. The key, he argues, is to replace reasoning by association with causal reasoning. Instead of the mere ability to correlate fever and malaria, machines need the capacity to reason that malaria causes fever. Once this kind of causal framework is in place, it becomes possible for machines to ask counterfactual questions — to inquire how the causal relationships would change given some kind of intervention — which Pearl views as the cornerstone of scientific thought. Pearl also proposes a formal language in which to make this kind of thinking possible — a 21st-century version of the Bayesian framework that allowed machines to think probabilistically.

Pearl expects that causal reasoning could provide machines with human-level intelligence. They'd be able to communicate with humans more effectively and even, he explains, achieve status as moral entities with a capacity for free will — and for evil. *Quanta Magazine* sat down with Pearl at a recent conference in San Diego and later held a follow-up interview with him by phone. An edited and condensed version of those conversations follows.

Why is your new book called “The Book of Why”?

It means to be a summary of the work I've been doing the past 25 years about cause and effect, what it means in one's life, its applications, and how we go about coming up with answers to questions that are inherently causal. Oddly, those questions have been abandoned by science. So I'm here to make up for the neglect of science.

That's a dramatic thing to say, that science has abandoned cause and effect. Isn't that exactly what all of science is about?

Of course, but you cannot see this noble aspiration in scientific equations. The language of algebra is symmetric: If X tells us about Y , then Y tells us about X . I'm talking about deterministic relationships. There's no way to write in mathematics a simple fact — for example, that the upcoming storm causes the barometer to go down, and not the other way around.

Mathematics has not developed the asymmetric language required to capture our understanding that if X causes Y that does not mean that Y causes X . It

sounds like a terrible thing to say against science, I know. If I were to say it to my mother, she'd slap me.

But science is more forgiving: Seeing that we lack a calculus for asymmetrical relations, science encourages us to create one. And this is where mathematics comes in. It turned out to be a great thrill for me to see that a simple calculus of causation solves problems that the greatest statisticians of our time deemed to be ill-defined or unsolvable. And all this with the ease and fun of finding a proof in high-school geometry.

You made your name in AI a few decades ago by teaching machines how to reason probabilistically. Explain what was going on in AI at the time.

The problems that emerged in the early 1980s were of a predictive or diagnostic nature. A doctor looks at a bunch of symptoms from a patient and wants to come up with the probability that the patient has malaria or some other disease. We wanted automatic systems, expert systems, to be able to replace the professional — whether a doctor, or an explorer for minerals, or some other kind of paid expert. So at that point I came up with the idea of doing it probabilistically.

Unfortunately, standard probability calculations required exponential space and exponential time. I came up with a scheme called Bayesian networks that required polynomial time and was also quite transparent.

Yet in your new book you describe yourself as an apostate in the AI community today. In what sense?

In the sense that as soon as we developed tools that enabled machines to reason with uncertainty, I left the arena to pursue a more challenging task: reasoning with cause and effect. Many of my AI colleagues are still occupied with uncertainty. There are circles of research that continue to work on diagnosis without worrying about the causal aspects of the problem. All they want is to predict well and to diagnose well.

I can give you an example. All the machine-learning work that we see today is conducted in diagnostic mode — say, labeling objects as “cat” or “tiger.” They

don't care about intervention; they just want to recognize an object and to predict how it's going to evolve in time.

I felt an apostate when I developed powerful tools for prediction and diagnosis knowing already that this is merely the tip of human intelligence. If we want machines to reason about interventions ("What if we ban cigarettes?") and introspection ("What if I had finished high school?"), we must invoke causal models. Associations are not enough — and this is a mathematical fact, not opinion.

People are excited about the possibilities for AI. You're not?

As much as I look into what's being done with deep learning, I see they're all stuck there on the level of associations. Curve fitting. That sounds like sacrilege, to say that all the impressive achievements of deep learning amount to just fitting a curve to data. From the point of view of the mathematical hierarchy, no matter how skillfully you manipulate the data and what you read into the data when you manipulate it, it's still a curve-fitting exercise, albeit complex and nontrivial.

The way you talk about curve fitting, it sounds like you're not very impressed with machine learning.

No, I'm very impressed, because we did not expect that so many problems could be solved by pure curve fitting. It turns out they can. But I'm asking about the future — what next? Can you have a robot scientist that would plan an experiment and find new answers to pending scientific questions? That's the next step. We also want to conduct some communication with a machine that is meaningful, and meaningful means matching our intuition. If you deprive the robot of your intuition about cause and effect, you're never going to communicate meaningfully. Robots could not say "I should have done better," as you and I do. And we thus lose an important channel of communication.

What are the prospects for having machines that share our intuition about cause and effect?

We have to equip machines with a model of the environment. If a machine does not have a model of reality, you cannot expect the machine to behave intelligently in that reality. The first step, one that will take place in maybe 10 years, is that conceptual models of reality will be programmed by humans.

The next step will be that machines will postulate such models on their own and will verify and refine them based on empirical evidence. That is what happened to science; we started with a geocentric model, with circles and epicycles, and ended up with a heliocentric model with its ellipses.

Robots, too, will communicate with each other and will translate this hypothetical world, this wild world, of metaphorical models.□

When you share these ideas with people working in AI today, how do they react?

AI is currently split. First, there are those who are intoxicated by the success of machine learning and deep learning and neural nets. They don't understand what I'm talking about. They want to continue to fit curves. But when you talk to people who have done any work in AI outside statistical learning, they get it immediately. I have read several papers written in the past two months about the limitations of machine learning.

Are you suggesting there's a trend developing away from machine learning?

Not a trend, but a serious soul-searching effort that involves asking: Where are we going? What's the next step?

That was the last thing I wanted to ask you.

I'm glad you didn't ask me about free will.

In that case, what do you think about free will?

We're going to have robots with free will, absolutely. We have to understand how to program them and what we gain out of it. For some reason, evolution has found this sensation of free will to be computationally desirable.

In what way?

You have the sensation of free will; evolution has equipped us with this sensation. Evidently, it serves some computational function.

Will it be obvious when robots have free will?

I think the first evidence will be if robots start communicating with each other counterfactually, like "You should have done better." If a team of robots playing soccer starts to communicate in this language, then we'll know that they have a sensation of free will. "You should have passed me the ball — I was waiting for you and you didn't!" "You should have" means you could have controlled whatever urges made you do what you did, and you didn't. So the first sign will be communication; the next will be better soccer.

Now that you've brought up free will, I guess I should ask you about the capacity for evil, which we generally think of as being contingent upon an ability to make choices. What is evil?

It's the belief that your greed or grievance supersedes all standard norms of society. For example, a person has something akin to a software module that says "You are hungry, therefore you have permission to act to satisfy your greed or grievance." But you have other software modules that instruct you to follow the standard laws of society. One of them is called compassion. When you elevate your grievance above those universal norms of society, that's evil.

So how will we know when AI is capable of committing evil?

When it is obvious for us that there are software components that the robot ignores, consistently ignores. When it appears that the robot follows the advice of some software components and not others, when the robot ignores the advice of other components that are maintaining norms of behavior that have been programmed into them or are expected to be there on the basis of past learning. And the robot stops following them.

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

COMPUTER SCIENCE

Smart Swarms Seek New Ways to Cooperate

New algorithms show how swarms of very simple robots can be made to work together as a group.

Individual flapping robots called “smarticles” can’t move on their own. But put them together and they can be programmed to move as a group.

Introduction

In a lab at the Georgia Institute of Technology, physicists run experiments with robots that look as though they came from the dollar store. The robots can’t move through space. They can’t communicate. Mostly they flap their little arms, like beetles stuck on their backs.

But put a lot of these objects together and you get something from nothing: They hit each other, nudge each other and tangle with each other. And eventually, they start to work as a unit.

Researchers are learning how to control these systems so that they function in a manner similar to swarms of bees or colonies of ants: Each individual operates in response to the same basic set of instructions. But when the swarm comes together, its members can carry out complex behaviors without any centralized direction.

“Our whole perspective is: What’s the simplest computational model that will achieve these complicated tasks?” said [Dana Randall](#)(opens a new tab), a computer scientist at Georgia Tech and one of the lead researchers on the project. “We’re looking for elegance and simplicity.”

As a computer scientist, Randall thinks about the problem in algorithmic terms: What is the most basic set of instructions individual elements in a swarm can run, based on the meager data they can collect, that will lead inevitably to the complex collective behavior researchers want? This past November, Randall and colleagues [published an algorithm](#)(opens a new tab) that ensures that an idealized particle swarm will move in a coordinated manner.

The work with these robots, known as “smarticles,” is part of a broader interest in the feasibility and applications of self-organizing robots. Other examples include [“droplet”-size robots](#) being developed at the University of Colorado, [“Kilobot” swarms](#) at Harvard University, and swarmanoids out of a [pioneering lab in Belgium](#). In many of these cases the idea is to mimic emergent phenomena found in nature, like the regimented motion of a decentralized colony of army ants or the unconscious, self-programming assembly of DNA molecules. “We know what we want the collective to do, but in order to program it we need to know what each agent must be doing on the individual level,” said [Melvin Gauci](#), a researcher at Harvard working on swarm robotics. “Going between those two levels is what’s very challenging.”

Dana Randall and Dan Goldman in Goldman’s lab at Georgia Tech.

Beware of Leaders

[Daniel Goldman](#) is a physicist at Georgia Tech who is leading the experiments with smarticles (a portmanteau of “smart active particles”). His fundamental scientific interest is in the physics of active granular materials that have the ability to change their own shape. In a slide deck he brings to conferences, he includes a clip from *Spider-Man 3* that shows [the birth of the supervillain Sandman](#): Loose grains of sand skitter across the desert and then congeal into the shape of a man. Smarticles are Goldman’s way of testing active granular materials in a lab.

“They give us a way to use geometry to control the properties of a material. If you blur your eyes, you can think of this collection of smarticles as a real material,” Goldman said.

The smarticles have short arms, like the prongs of a staple, that they can swing back and forth. They respond to light and tones of different frequencies. They can also be programmed to adjust the rate at which they swing their arms in response to the other smarticles they encounter in their immediate vicinity.

There are a few basic maneuvers you might like a swarm of smarticles to perform: compression (packing together), expansion (spreading out) and locomotion. These maneuvers could serve as building blocks for more complicated feats, but even the most basic functions, like compression, are

hard to engineer when none of the smarticles have any idea where they're positioned in relation to the overall group.

To understand the opportunity and the challenge of engineering complex behavior from simple parts, it's worth keeping in mind what an individual smarticle knows. And the answer is: not much. It can't see, it has limited memory, and the only thing it knows about the other smarticles it's supposed to coordinate with is what it can learn from bumping into its immediate neighbors.

"Imagine one person at a rock concert with his eyes closed," said Joshua Daymude, a graduate student in computer science at Arizona State University who works on the smarticles project.

Video: Goldman explains how "smarticles" work together to demonstrate collective behavior.

[Ben Rollins](#)[\(opens a new tab\)](#) for Quanta Magazine

One strategy would be to appoint a leader that orchestrates the swarm, but that approach is vulnerable to disruption — if the leader goes down, the whole swarm goes down. Another is to give each robot in the swarm a unique job to perform, but that's impractical to implement on a large scale. "Individually programming 1,000 robots is basically an impossible task," said [Jeff Dusek](#)[\(opens a new tab\)](#), a researcher at Olin College of Engineering and a former member of the Self-Organizing Systems research group at Harvard, where he worked on [underwater robot swarms](#)[\(opens a new tab\)](#). But "if every agent is following the same set of rules, your code is exactly the same whether you have 10 or 1,000 or 10,000 agents."

An algorithm used to program a swarm has two properties. First, it's distributed, meaning it runs separately on each individual particle in the system (the way each army ant carries out the same simple set of instructions based on whatever it senses about its local environment). Second, it incorporates randomness. This means that if an army ant senses, say, five other army ants right around it, maybe there's a 20 percent chance it moves to the left and an 80 percent chance it moves to the right. Randomized algorithms contrast with deterministic algorithms, where each step is completely determined by what came before it.

Randomness might seem undesirable in an algorithm — after all, when you implement a procedure you'd typically like certainty about the outcome. But randomness also conveys some surprising performance advantages that, among other things, make randomized algorithms well-suited for application to smarticle swarms.

Random Guarantees

In 2015, Goldman and Randall were discussing the possibility of finding rules that would lead Goldman's smarticles to act coherently as a group. Randall realized that the swarm behaviors Goldman was after were similar to the behavior of idealized particle systems studied in computer science.

"I was like, 'I know exactly what's going on,'" Randall said.

For Randall, the smarticles' behavior resembled emergent phenomena modeled by computer scientists in many other contexts. One of the most famous examples is how segregated neighborhoods form. In the late 1960s, the economist Thomas Schelling wanted to understand how housing segregation takes hold in the absence of any centralized power sorting people into neighborhoods by skin color. He imagined a hypothetical person looking at his neighbors and deciding to move elsewhere based on how many of them looked like him. When the person moved, Schelling transported him to a random spot in the housing grid where he repeated the algorithmic process of observing his neighbors and deciding whether to stay or go. Schelling discovered that, according to his rules, residential segregation is virtually guaranteed to take hold, even if individuals prefer to live in diverse neighborhoods.

William Savoie, a doctoral student in Goldman's lab.

Randall realized that smarticles in a swarm resemble people in the Schelling model. In both cases, individuals have to make decisions without knowing their global position (they only know what they can see around them). And in Schelling's model the decisions can be made with an element of randomness — if your neighbors look different from you, maybe there's a high probability you move, but also some small probability you choose to stay put.

In 2016, Randall and her collaborators [published a paper](#) (opens a new tab) that imagined idealized particles living on a grid and deciding whether to move or stay put based on how many other particles they observed around them. The decision making was probabilistic — particles would essentially “roll” a weighted die every time they had to make a choice. Randall and her co-authors proved that if they weighted the die correctly, they were guaranteed to end up with a compressed swarm (in the same way Schelling could have proved that if he set individuals’ tolerance for diversity at the right level, segregation was unavoidable). By adjusting parameters in the algorithm, they could also guarantee that a particle swarm would move to an expanded state.

The randomness in the algorithm helps particles in a swarm avoid getting stuck in locally compressed states, where lots of isolated subgroups are clustered together but the swarm as a whole isn’t compressed. The randomness ensures that if smarticles end up in small compressed groups, there’s a chance individuals will still decide to move to a new location, keeping the process alive until an overall compressed state is reached. (It takes just a little randomness to nudge particles out of locally compressed states — it takes a lot more to nudge them out of a globally compressed state.)

Into the World

Proving that particles in a theoretical world can run a simple algorithm and achieve specific swarm behaviors is one thing. Actually implementing the algorithm in cheap, faulty, real-life smarticles clanking around in a box is another.

“Our theory collaborators are coming up with ways to program these things, but we’re just in the beginning and we can’t yet say these schemes have been transferred directly,” Goldman said.

One problem has been to get the smarticles to move as a group. At first, if the researchers packed smarticles into a confined space, the ensemble would just stagger around randomly.

But one day the physicists were observing this chaotic motion when the battery died in one of the smarticles. Goldman and his collaborators noticed that the swarm suddenly started moving in the direction of the inactive unit. The researchers reported this inadvertent finding to their computer science

collaborators, who seized on the clue. The work led to the recent development of the algorithm that will always get an idealized swarm to move in a specified direction.

Little by little, the computer science experiments and the physical ones are moving closer together. The researchers hope to eventually prove theoretically that a basic algorithm, implemented in a distributed way in a large collection of small, cheap robots, is guaranteed to produce a specified swarm behavior.

“We’d like to move to a point where it’s not that batteries died and we found a phenomenon,” Daymude said. “We’d like it to be more intentional.”

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

ARTIFICIAL INTELLIGENCE

Machine Learning Confronts the Elephant in the Room

A visual prank exposes an Achilles’ heel of computer vision systems: Unlike humans, they can’t do a double take.

Introduction

Score one for the human brain. In [a new study](#)(opens a new tab), computer scientists found that artificial intelligence systems fail a vision test a child could accomplish with ease.

“It’s a clever and important study that reminds us that ‘deep learning’ isn’t really that deep,” said [Gary Marcus](#)(opens a new tab), a neuroscientist at New York University who was not affiliated with the work.

The result takes place in the field of computer vision, where artificial intelligence systems attempt to detect and categorize objects. They might try to find all the pedestrians in a street scene, or just distinguish a bird from a bicycle (which is a notoriously difficult task). The stakes are high: As computers take over critical tasks like automated surveillance and autonomous driving, we’ll want their visual processing to be at least as good as the human eyes they’re replacing.

It won't be easy. The new work accentuates the sophistication of human vision — and the challenge of building systems that mimic it. In the study, the researchers presented a computer vision system with a living room scene. The system processed it well. It correctly identified a chair, a person, books on a shelf. Then the researchers introduced an anomalous object into the scene — an image of an elephant. The elephant's mere presence caused the system to forget itself: Suddenly it started calling a chair a couch and the elephant a chair, while turning completely blind to other objects it had previously seen.

“There are all sorts of weird things happening that show how brittle current object detection systems are,” said [Amir Rosenfeld](#) (opens a new tab), a researcher at York University in Toronto and co-author of the study along with his York colleague [John Tsotsos](#) (opens a new tab) and [Richard Zemel](#) (opens a new tab) of the University of Toronto.

Researchers are still trying to understand exactly why computer vision systems get tripped up so easily, but they have a good guess. It has to do with an ability humans have that AI lacks: the ability to understand when a scene is confusing and thus go back for a second glance.

The Elephant in the Room

Eyes wide open, we take in staggering amounts of visual information. The human brain processes it in stride. “We open our eyes and everything happens,” said Tsotsos.

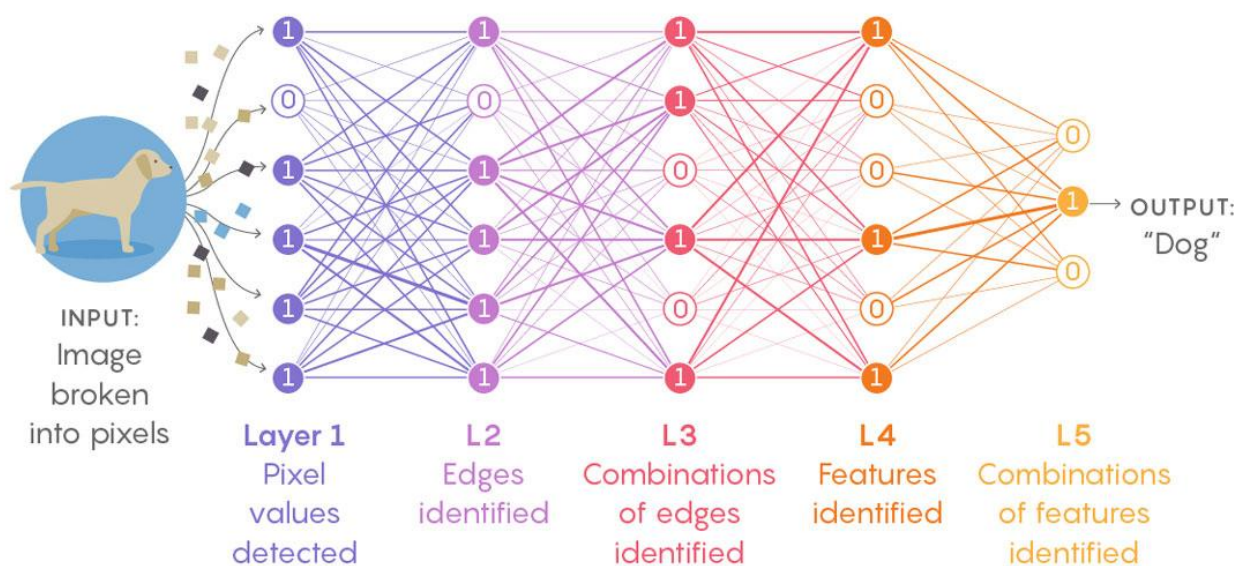
Artificial intelligence, by contrast, creates visual impressions laboriously, as if it were reading a description in Braille. It runs its algorithmic fingertips over pixels, which it shapes into increasingly complex representations. The specific type of AI system that performs this process is called a neural network. It sends an image through a series of “layers.” At each layer, the details of the image — the colors and brightnesses of individual pixels — give way to increasingly

abstracted descriptions of what the image portrays. At the end of the process, the neural network produces a best-guess prediction about what it's looking at.

“It's all moving from one layer to the next by taking the output of the previous layer, processing it and passing it along to the next layer, like a pipeline,” said Tsotsos.

Learning From Experience

Deep neural networks learn by adjusting the strengths of their connections to better convey input signals through multiple layers to neurons associated with the right general concepts.



When data is fed into a network, each artificial neuron that fires (labeled "1") transmits signals to certain neurons in the next layer, which are likely to fire if multiple signals are received. The process filters out noise and retains only the most relevant features.

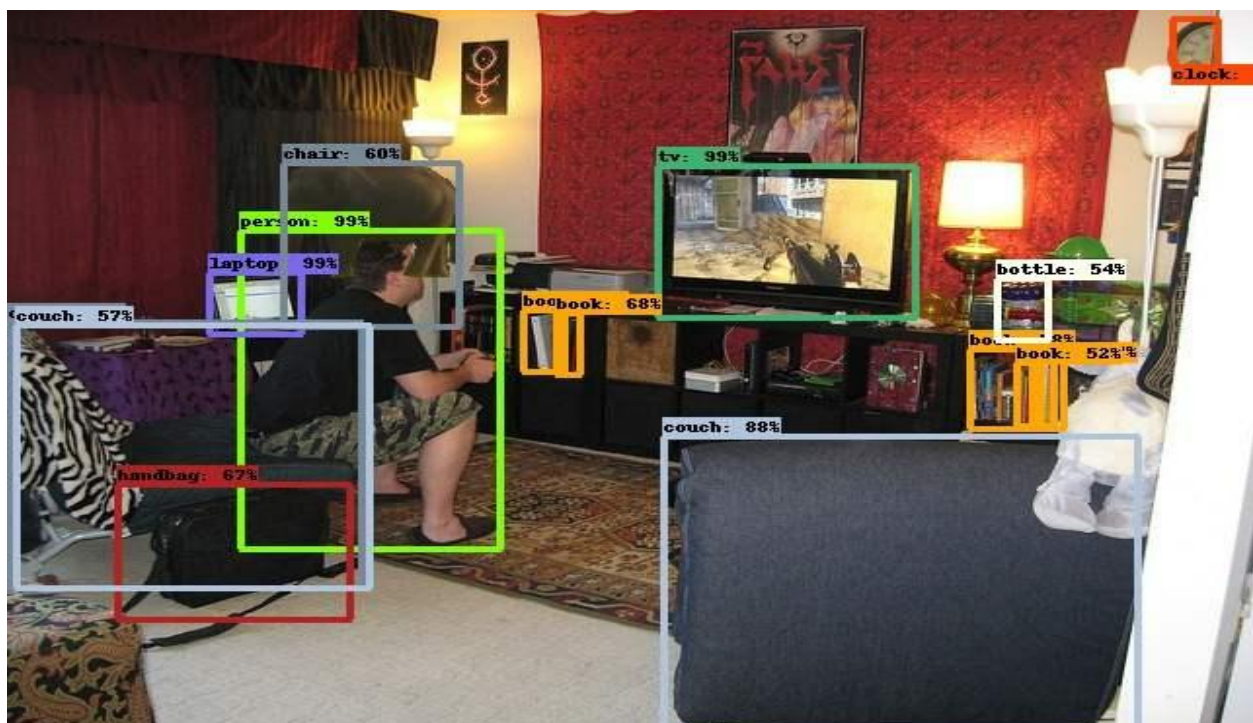
Lucy Reading-Ikkanda/Quanta Magazine

Neural networks are adept at specific visual chores. They can outperform humans in narrow tasks like sorting objects into best-fit categories — labeling

dogs with their breed, for example. These successes have raised expectations that computer vision systems might soon be good enough to steer a car through crowded city streets.

They've also provoked researchers to probe their vulnerabilities. In recent years there have been a slew of attempts, known as "adversarial attacks," in which researchers contrive scenes to make neural networks fail. In one experiment, computer scientists tricked a neural network into [mistaking a turtle for a rifle](#)(opens a new tab). In another, researchers waylaid a neural network by placing an [image of a psychedelically colored toaster](#)(opens a new tab) alongside ordinary objects like a banana.

This new study has the same spirit. The three researchers fed a neural network a living room scene: A man seated on the edge of a shabby chair leans forward as he plays a video game. After chewing on this scene, a neural network correctly detected a number of objects with high confidence: a person, a couch, a television, a chair, some books.



In the unmodified image at left, the neural network correctly identifies many items in a cluttered living room scene with high probability. Add an elephant, as in the image at right, and problems arise. The chair in the lower-left corner becomes a couch, the nearby cup disappears, and the elephant gets misidentified as a chair.

Then the researchers introduced something incongruous into the scene: an image of an elephant in semiprofile. The neural network started getting its pixels crossed. In some trials, the elephant led the neural network to misidentify the chair as a couch. In others, the system overlooked objects, like a row of books, that it had correctly detected in earlier trials. These errors occurred even when the elephant was far from the mistaken objects.

Snafus like those extrapolate in unsettling ways to autonomous driving. A computer can't drive a car if it might go blind to a pedestrian just because a second earlier it passed a turkey on the side of the road.

And as for the elephant itself, the neural network was all over the place: Sometimes the system identified it correctly, sometimes it called the elephant a sheep, and sometimes it overlooked the elephant completely.

"If there is actually an elephant in the room, you as a human would likely notice it," said Rosenfeld. "The system didn't even detect its presence."

Everything Connected to Everything

When human beings see something unexpected, we do a double take. It's a common phrase with real cognitive implications — and it explains why neural networks fail when scenes get weird.

Today's best neural networks for object detection work in a "feed forward" manner. This means that information flows through them in only one direction. They start with an input of fine-grained pixels, then move to curves, shapes, and scenes, with the network making its best guess about what it's seeing at each step along the way. As a consequence, errant observations early in the process end up contaminating the end of the process, when the neural network pools together everything it thinks it knows in order to make a guess about what it's looking at.

Amir Rosenfeld, a researcher at York University in Toronto, said that the new work shows "how brittle current object detection systems are."
Courtesy of Amir Rosenfeld

“By the top of the neural network you have everything connected to everything, so you have the potential to have every feature in every location interfering with every possible output,” said Tsotsos.

The human way is better. Imagine you’re given a very brief glimpse of an image containing a circle and a square, with one of them colored blue and the other red. Afterward you’re asked to name the color of the square. With only a single glance to go on, you’re likely to confuse the colors of the two shapes. But you’re also likely to recognize that you’re confused and to ask for another look. And, critically, when you take that second look, you know to focus your attention on just the color of the square.

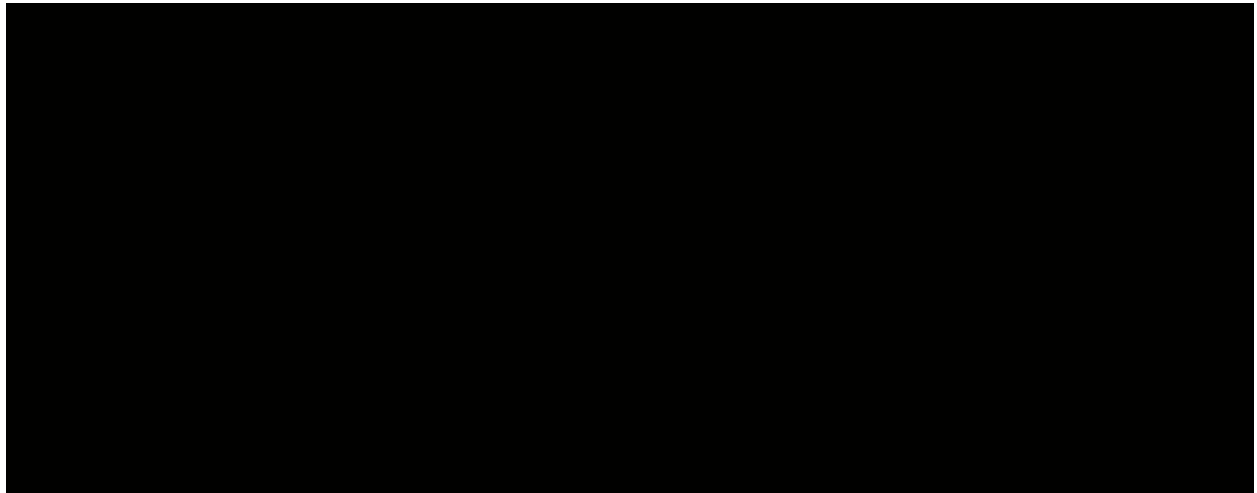
“The human visual system says, ‘I don’t have right answer yet, so I have to go backwards to see where I might have made an error,’” explained Tsotsos, who has been developing a theory called [selective tuning](#)[\(opens a new tab\)](#) that explains this feature of visual cognition.

Most neural networks lack this ability to go backward. It’s a hard trait to engineer. One advantage of feed-forward networks is that they’re relatively straightforward to train — process an image through these six layers and get an answer. But if neural networks are to have license to do a double take, they’ll need a sophisticated understanding of when to draw on this new capacity (when to look twice) and when to plow ahead in a feed-forward way. Human brains switch between these different processes seamlessly; neural networks will need a new theoretical framework before they can do the same.

Leading researchers in the world are working on it, though, and they’re calling for backup. Earlier this month, Google AI announced a contest to [crowdsource image classifiers](#)[\(opens a new tab\)](#) that can see their way through adversarial attacks. The winning entry will need to unambiguously distinguish between an image of a bird and an image of a bicycle. It would be a modest first step — but also a necessary one.

Machine Learning's 'Amazing' Ability to Predict Chaos

In new computer experiments, artificial-intelligence algorithms can tell the future of chaotic systems.



Researchers have used machine learning to predict the chaotic evolution of a model flame front.

Introduction

Half a century ago, the pioneers of [chaos theory](#) discovered that the “butterfly effect” makes long-term prediction impossible. Even the smallest perturbation to a complex system (like the weather, the economy or just about anything else) can touch off a concatenation of events that leads to a dramatically divergent future. Unable to pin down the state of these systems precisely enough to predict how they’ll play out, we live under a veil of uncertainty.

But now the robots are here to help.

In a series of results reported in the journals [Physical Review Letters](#) ([opens a new tab](#)) and [Chaos](#) ([opens a new tab](#)), scientists have used [machine learning](#) — the same computational technique behind recent successes in artificial intelligence — to predict the future evolution of chaotic systems out to

stunningly distant horizons. The approach is being lauded by outside experts as groundbreaking and likely to find wide application.

“I find it really amazing how far into the future they predict” a system’s chaotic evolution, said [Herbert Jaeger](#)(opens a new tab), a professor of computational science at Jacobs University in Bremen, Germany.

The findings come from veteran chaos theorist [Edward Ott](#)(opens a new tab) and four collaborators at the University of Maryland. They employed a machine-learning algorithm called reservoir computing to “learn” the dynamics of an archetypal chaotic system called the Kuramoto-Sivashinsky equation. The evolving solution to this equation behaves like a flame front, flickering as it advances through a combustible medium. The equation also describes drift waves in plasmas and other phenomena, and serves as “a test bed for studying turbulence and spatiotemporal chaos,” said [Jaideep Pathak](#)(opens a new tab), Ott’s graduate student and the lead author of the new papers.

Jaideep Pathak, Michelle Girvan, Brian Hunt and Edward Ott of the University of Maryland, who (along with Zhixin Lu, now of the University of Pennsylvania) have shown that machine learning is a powerful tool for predicting chaos.

Faye Levine/University of Maryland

After training itself on data from the past evolution of the Kuramoto-Sivashinsky equation, the researchers’ reservoir computer could then closely predict how the flamelike system would continue to evolve out to eight “Lyapunov times” into the future, eight times further ahead than previous methods allowed, loosely speaking. The Lyapunov time represents how long it takes for two almost-identical states of a chaotic system to exponentially diverge. As such, it typically sets the horizon of predictability.

“This is really very good,” [Holger Kantz](#)(opens a new tab), a chaos theorist at the Max Planck Institute for the Physics of Complex Systems in Dresden, Germany, said of the eight-Lyapunov-time prediction. “The machine-learning technique is almost as good as knowing the truth, so to say.”

The algorithm knows nothing about the Kuramoto-Sivashinsky equation itself; it only sees data recorded about the evolving solution to the equation. This makes the machine-learning approach powerful; in many cases, the equations describing a chaotic system aren’t known, crippling dynamicists’ efforts to model and predict them. Ott and company’s results suggest you don’t

need the equations — only data. “This paper suggests that one day we might be able perhaps to predict weather by machine-learning algorithms and not by sophisticated models of the atmosphere,” Kantz said.

Besides weather forecasting, experts say the machine-learning technique could help with monitoring cardiac arrhythmias for signs of impending heart attacks and monitoring neuronal firing patterns in the brain for signs of neuron spikes. More speculatively, it might also help with predicting rogue waves, which endanger ships, and possibly even earthquakes.

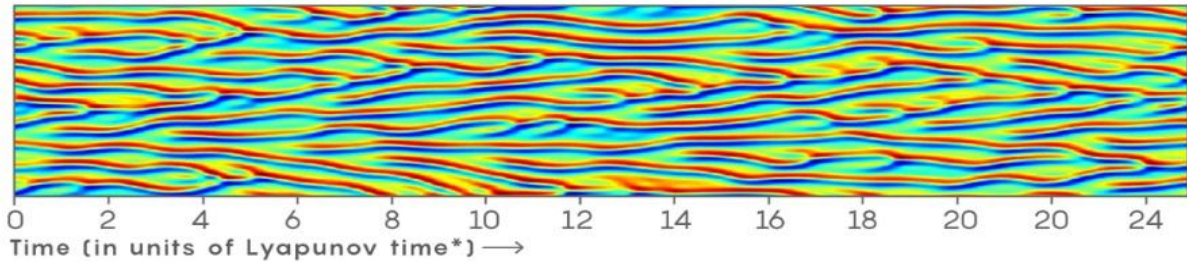
Ott particularly hopes the new tools will prove useful for giving advance warning of solar storms, like the one that erupted across 35,000 miles of the sun’s surface in 1859. That magnetic outburst created aurora borealis visible all around the Earth and blew out some telegraph systems, while generating enough voltage to allow other lines to operate with their power switched off. If such a solar storm lashed the planet unexpectedly today, experts say it would severely damage Earth’s electronic infrastructure. “If you knew the storm was coming, you could just turn off the power and turn it back on later,” Ott said.

Training Computers to Tame Chaos

A machine-learning algorithm has been shown to accurately predict a chaotic system far further into the future than previously possible.

A Chaos Model

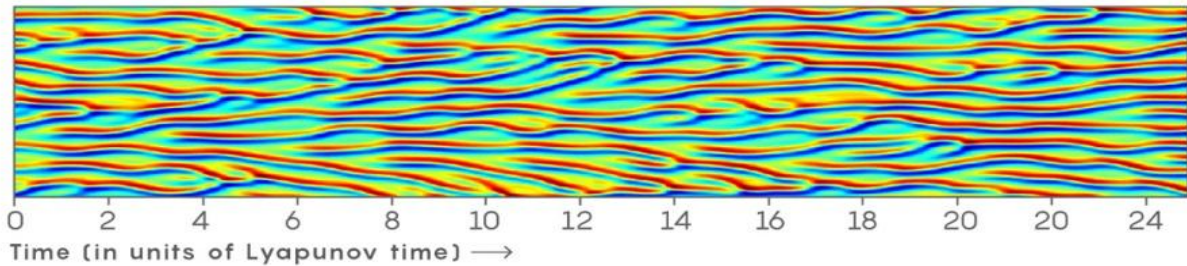
Researchers started with the evolving solution to the Kuramoto-Sivashinsky equation, which models propagating flames:



* Lyapunov time = Length of time before a small difference in the system's initial state begins to diverge exponentially. It typically sets the horizon of predictability, which varies from system to system.

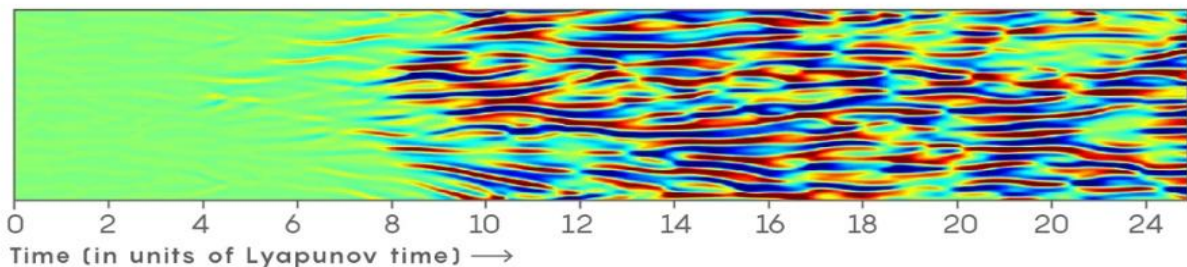
B Machine Learning

After training itself on data from the past evolution of the Kuramoto-Sivashinsky system, the "reservoir computing" algorithm predicts its future evolution:



A - B Do They Match?

Subtracting B from A shows that the algorithm accurately predicts the model out to an impressive 8 Lyapunov times, before chaos ultimately prevails:



He, Pathak and their colleagues [Brian Hunt](#)(opens a new tab), [Michelle Girvan](#)(opens a new tab) and [Zhixin Lu](#)(opens a new tab) (who is now at the University of Pennsylvania) achieved their results by synthesizing existing tools. Six or seven years ago, when the powerful algorithm known as “deep learning” was starting to master AI tasks like image and speech recognition, they started reading up on machine learning and thinking of clever ways to apply it to chaos. They learned of a handful of promising results predating the deep-learning revolution. Most importantly, in the early 2000s, Jaeger and fellow German chaos theorist Harald Haas [made use](#)(opens a new tab) of a network of randomly connected artificial neurons — which form the “reservoir” in reservoir computing — to learn the dynamics of three chaotically coevolving variables. After training on the three series of numbers, the network could predict the future values of the three variables out to an impressively distant horizon. However, when there were more than a few interacting variables, the computations became impossibly unwieldy. Ott and his colleagues needed a more efficient scheme to make reservoir computing relevant for large chaotic systems, which have huge numbers of interrelated variables. Every position along the front of an advancing flame, for example, has velocity components in three spatial directions to keep track of.

It took years to strike upon the straightforward solution. “What we exploited was the locality of the interactions” in spatially extended chaotic systems, Pathak said. Locality means variables in one place are influenced by variables at nearby places but not by places far away. “By using that,” Pathak explained, “we can essentially break up the problem into chunks.” That is, you can parallelize the problem, using one reservoir of neurons to learn about one patch of a system, another reservoir to learn about the next patch, and so on, with slight overlaps of neighboring domains to account for their interactions.

Parallelization allows the reservoir computing approach to handle chaotic systems of almost any size, as long as proportionate computer resources are dedicated to the task.

If we have ignorance we should use the machine learning to fill in the gaps where the ignorance resides.

[Edward Ott](#)

Ott explained reservoir computing as a three-step procedure. Say you want to use it to predict the evolution of a spreading fire. First, you measure the height of the flame at five different points along the flame front, continuing to measure the height at these points on the front as the flickering flame advances over a period of time. You feed these data-streams in to randomly chosen artificial neurons in the reservoir. The input data triggers the neurons to fire, triggering connected neurons in turn and sending a cascade of signals throughout the network.

The second step is to make the neural network learn the dynamics of the evolving flame front from the input data. To do this, as you feed data in, you also monitor the signal strengths of several randomly chosen neurons in the reservoir. Weighting and combining these signals in five different ways produces five numbers as outputs. The goal is to adjust the weights of the various signals that go into calculating the outputs until those outputs consistently match the next set of inputs — the five new heights measured a moment later along the flame front. “What you want is that the output should be the input at a slightly later time,” Ott explained.

To learn the correct weights, the algorithm simply compares each set of outputs, or predicted flame heights at each of the five points, to the next set of inputs, or actual flame heights, increasing or decreasing the weights of the various signals each time in whichever way would have made their combinations give the correct values for the five outputs. From one time-step to the next, as the weights are tuned, the predictions gradually improve, until the algorithm is consistently able to predict the flame’s state one time-step later.

“In the third step, you actually do the prediction,” Ott said. The reservoir, having learned the system’s dynamics, can reveal how it will evolve. The network essentially asks itself what will happen. Outputs are fed back in as the new inputs, whose outputs are fed back in as inputs, and so on, making a projection of how the heights at the five positions on the flame front will evolve. Other reservoirs working in parallel predict the evolution of height elsewhere in the flame.

In a plot in their [PRL](#) paper, which appeared in January, the researchers show that their predicted flamelike solution to the Kuramoto-Sivashinsky equation exactly matches the true solution out to eight Lyapunov times before chaos finally wins, and the actual and predicted states of the system diverge.

The usual approach to predicting a chaotic system is to measure its conditions at one moment as accurately as possible, use these data to calibrate a physical model, and then evolve the model forward. As a ballpark estimate, you'd have to measure a typical system's initial conditions 100,000,000 times more accurately to predict its future evolution eight times further ahead.

The machine-learning technique is almost as good as knowing the truth.

Holger Kantz

That's why machine learning is "a very useful and powerful approach," said [Ulrich Parlitz](#) ([opens a new tab](#)) of the Max Planck Institute for Dynamics and Self-Organization in Göttingen, Germany, who, like Jaeger, also applied machine learning to low-dimensional chaotic systems in the early 2000s. "I think it's not only working in the example they present but is universal in some sense and can be applied to many processes and systems." In [a paper soon to be published in *Chaos*](#) ([opens a new tab](#)), Parlitz and a collaborator applied reservoir computing to predict the dynamics of "excitable media," such as cardiac tissue. Parlitz suspects that deep learning, while being more complicated and computationally intensive than reservoir computing, will also work well for tackling chaos, as will other machine-learning algorithms. Recently, researchers at the Massachusetts Institute of Technology and ETH Zurich [achieved similar results](#) ([opens a new tab](#)) as the Maryland team using a "long short-term memory" neural network, which has recurrent loops that enable it to store temporary information for a long time. Since the work in their [PRL](#) paper, Ott, Pathak, Girvan, Lu and other collaborators have come closer to a practical implementation of their prediction technique. In [new research accepted for publication in *Chaos*](#) ([opens a new tab](#)), they showed that improved predictions of chaotic systems like the Kuramoto-Sivashinsky equation become possible by hybridizing the data-driven, machine-learning approach and traditional model-based prediction. Ott sees this as a more likely avenue for improving weather prediction and similar efforts, since we don't always have complete high-resolution data or perfect physical models. "What we should do is use the good knowledge that we have where we have it," he said, "and if we have ignorance we should use the machine learning to fill in the gaps where the ignorance resides." The reservoir's predictions can essentially calibrate the models; in the case of the Kuramoto-Sivashinsky equation, accurate predictions are extended out to 12 Lyapunov times.

The duration of a Lyapunov time varies for different systems, from milliseconds to millions of years. (It's a few days in the case of the weather.)

The shorter it is, the touchier or more prone to the butterfly effect a system is, with similar states departing more rapidly for disparate futures. Chaotic systems are everywhere in nature, going haywire more or less quickly. Yet strangely, chaos itself is hard to pin down. “It’s a term that most people in dynamical systems use, but they kind of hold their noses while using it,” said [Amie Wilkinson](#)(opens a new tab), a professor of mathematics at the University of Chicago. “You feel a bit cheesy for saying something is chaotic,” she said, because it grabs people’s attention while having no agreed-upon mathematical definition or necessary and sufficient conditions. “There is no easy concept,” Kantz agreed. In some cases, tuning a single parameter of a system can make it go from chaotic to stable or vice versa.

Wilkinson and Kantz both define chaos in terms of stretching and folding, much like the repeated stretching and folding of dough in the making of puff pastries. Each patch of dough stretches horizontally under the rolling pin, separating exponentially quickly in two spatial directions. Then the dough is folded and flattened, compressing nearby patches in the vertical direction. The weather, wildfires, the stormy surface of the sun and all other chaotic systems act just this way, Kantz said. “In order to have this exponential divergence of trajectories you need this stretching, and in order not to run away to infinity you need some folding,” where folding comes from nonlinear relationships between variables in the systems.

The stretching and compressing in the different dimensions correspond to a system’s positive and negative “Lyapunov exponents,” respectively. In [another recent paper in *Chaos*](#)(opens a new tab), the Maryland team reported that their reservoir computer could successfully learn the values of these characterizing exponents from data about a system’s evolution. Exactly why reservoir computing is so good at learning the dynamics of chaotic systems is not yet well understood, beyond the idea that the computer tunes its own formulas in response to data until the formulas replicate the system’s dynamics. The technique works so well, in fact, that Ott and some of the other Maryland researchers now intend to use chaos theory as a way to better understand the internal machinations of neural networks.

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

New Theory Cracks Open the Black Box of Deep Learning

A new idea called the “information bottleneck” is helping to explain the puzzling success of today’s artificial-intelligence algorithms — and might also explain how human brains learn.

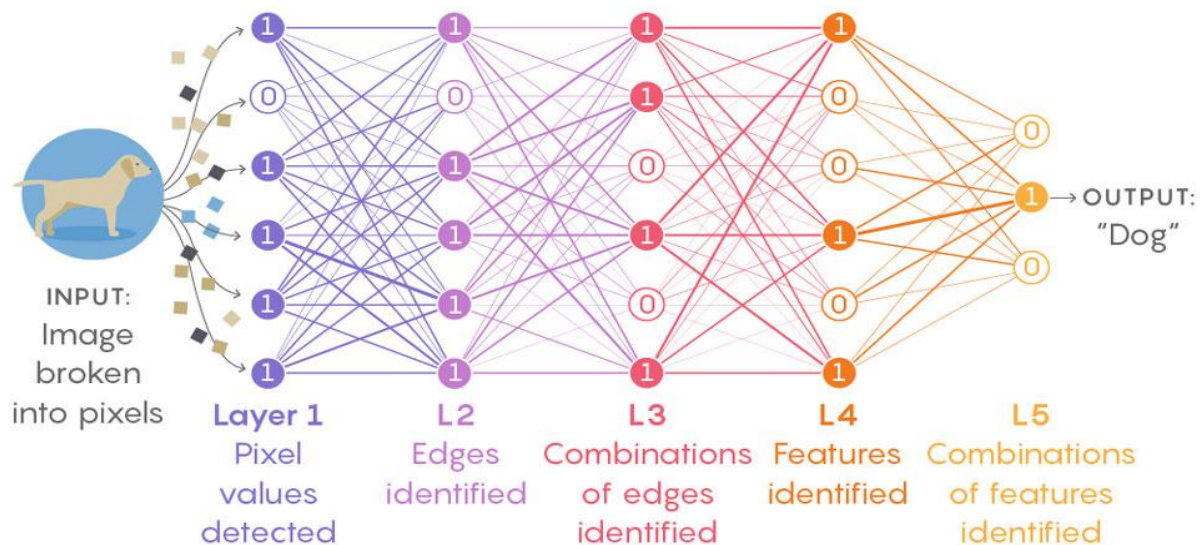
Introduction

Even as machines known as “deep neural networks” have learned to converse, drive cars, [beat video games](#) and [Go champions](#), dream, paint pictures and help make scientific discoveries, they have also confounded their human creators, who never expected so-called “deep-learning” algorithms to work so well. No underlying principle has guided the design of these learning systems, other than vague inspiration drawn from the architecture of the brain (and no one really understands how that operates either).

Like a brain, a deep neural network has layers of neurons — artificial ones that are figments of computer memory. When a neuron fires, it sends signals to connected neurons in the layer above. During deep learning, connections in the network are strengthened or weakened as needed to make the system better at sending signals from input data — the pixels of a photo of a dog, for instance — up through the layers to neurons associated with the right high-level concepts, such as “dog.” After a deep neural network has “learned” from thousands of sample dog photos, it can identify dogs in new photos as accurately as people can. The magic leap from special cases to general concepts during learning gives deep neural networks their power, just as it underlies human reasoning, creativity and the other faculties collectively termed “intelligence.” Experts wonder what it is about deep learning that enables generalization — and to what extent brains apprehend reality in the same way.

Learning From Experience

Deep neural networks learn by adjusting the strengths of their connections to better convey input signals through multiple layers to neurons associated with the right general concepts.



When data is fed into a network, each artificial neuron that fires (labeled "1") transmits signals to certain neurons in the next layer, which are likely to fire if multiple signals are received. The process filters out noise and retains only the most relevant features.

Last month, a [YouTube video\(opens a new tab\)](#) of a conference talk in Berlin, shared widely among artificial-intelligence researchers, offered a possible answer. In the talk, [Naftali Tishby\(opens a new tab\)](#), a computer scientist and neuroscientist from the Hebrew University of Jerusalem, presented evidence in support of a new theory explaining how deep learning works. Tishby argues that deep neural networks learn according to a procedure called the "information bottleneck," which he and two collaborators [first described in purely theoretical terms in 1999\(opens a new tab\)](#). The idea is that a network rids noisy input data of extraneous details as if by squeezing the information through a bottleneck, retaining only the features most relevant to general concepts. Striking new [computer experiments\(opens a new tab\)](#) by Tishby and

his student Ravid Shwartz-Ziv reveal how this squeezing procedure happens during deep learning, at least in the cases they studied. Tishby's findings have the AI community buzzing. "I believe that the information bottleneck idea could be very important in future deep neural network research," said [Alex Alemi](#) (opens a new tab) of Google Research, who has already [developed new approximation methods](#) (opens a new tab) for applying an information bottleneck analysis to large deep neural networks. The bottleneck could serve "not only as a theoretical tool for understanding why our neural networks work as well as they do currently, but also as a tool for constructing new objectives and architectures of networks," Alemi said. Some researchers remain skeptical that the theory fully accounts for the success of deep learning, but [Kyle Cranmer](#) (opens a new tab), a particle physicist at New York University who uses machine learning to analyze particle collisions at the Large Hadron Collider, said that as a general principle of learning, it "somehow smells right."

[Geoffrey Hinton](#) (opens a new tab), a pioneer of deep learning who works at Google and the University of Toronto, emailed Tishby after watching his Berlin talk. "It's extremely interesting," Hinton wrote. "I have to listen to it another 10,000 times to really understand it, but it's very rare nowadays to hear a talk with a really original idea in it that may be the answer to a really major puzzle."

According to Tishby, who views the information bottleneck as a fundamental principle behind learning, whether you're an algorithm, a housefly, a conscious being, or a physics calculation of emergent behavior, that long-awaited answer "is that the most important part of learning is actually forgetting."

The Bottleneck

Tishby began contemplating the information bottleneck around the time that other researchers were first mulling over deep neural networks, though neither concept had been named yet. It was the 1980s, and Tishby was thinking about how good humans are at speech recognition — a major challenge for AI at the time. Tishby realized that the crux of the issue was the question of relevance: What are the most relevant features of a spoken word, and how do we tease these out from the variables that accompany them, such as accents, mumbling and intonation? In general, when we face the sea of data that is reality, which signals do we keep?

Naftali Tishby, a professor of computer science at the Hebrew University of Jerusalem.

Miriam Alster, Flash 90. ELSC Art and Brain Week 2016

“This notion of relevant information was mentioned many times in history but never formulated correctly,” Tishby said in an interview last month. “For many years people thought information theory wasn’t the right way to think about relevance, starting with misconceptions that go all the way to Shannon himself.”

Claude Shannon, the founder of information theory, in a sense liberated the study of information starting in the 1940s by allowing it to be considered in the abstract — as 1s and 0s with purely mathematical meaning. Shannon took the view that, as Tishby put it, “information is not about semantics.” But, Tishby argued, this isn’t true. Using information theory, he realized, “you can define ‘relevant’ in a precise sense.”

Imagine $\mathbf{X}$ is a complex data set, like the pixels of a dog photo, and $\mathbf{Y}$ is a simpler variable represented by those data, like the word “dog.” You can capture all the “relevant” information in $\mathbf{X}$ about $\mathbf{Y}$ by compressing $\mathbf{X}$ as much as you can without losing the ability to predict $\mathbf{Y}$. In their 1999 paper, Tishby and co-authors [Fernando Pereira](#)(opens a new tab), now at Google, and [William Bialek](#)(opens a new tab), now at Princeton University, formulated this as a mathematical optimization problem. It was a fundamental idea with no killer application.

“I’ve been thinking along these lines in various contexts for 30 years,” Tishby said. “My only luck was that deep neural networks became so important.”

Eyeballs on Faces on People on Scenes

Though the concept behind deep neural networks had been kicked around for decades, their performance in tasks like speech and image recognition only took off in the early 2010s, due to improved training regimens and more powerful computer processors. Tishby recognized their potential connection to the information bottleneck principle in 2014 after reading a [surprising paper](#)(opens a new tab) by the physicists [David Schwab](#)(opens a new tab) and [Pankaj Mehta](#)(opens a new tab).

The [duo discovered](#) that a deep-learning algorithm invented by Hinton called the “deep belief net” works, in a particular case, exactly like renormalization, a technique used in physics to zoom out on a physical system by coarse-graining over its details and calculating its overall state. When Schwab and Mehta applied the deep belief net to a model of a magnet at its “critical point,” where the system is fractal, or self-similar at every scale, they found that the network automatically used the renormalization-like procedure to discover the model’s state. It was a stunning indication that, as the biophysicist [Ilya Nemenman](#)[\(opens a new tab\)](#) said at the time, “extracting relevant features in the context of statistical physics and extracting relevant features in the context of deep learning are not just similar words, they are one and the same.”

The only problem is that, in general, the real world isn’t fractal. “The natural world is not ears on ears on ears on ears; it’s eyeballs on faces on people on scenes,” Cranmer said. “So I wouldn’t say [the renormalization procedure] is why deep learning on natural images is working so well.” But Tishby, who at the time was undergoing chemotherapy for pancreatic cancer, realized that both deep learning and the coarse-graining procedure could be encompassed by a broader idea. “Thinking about science and about the role of my old ideas was an important part of my healing and recovery,” he said.

Noga Zaslavsky, left, and Ravid Shwartz-Ziv helped develop the information bottleneck theory of deep learning as graduate students of Naftali Tishby’s. Yifat Yogev (left); courtesy of Ravid Shwartz-Ziv (right)

In 2015, he and his student Noga Zaslavsky [hypothesized](#)[\(opens a new tab\)](#) that deep learning is an information bottleneck procedure that compresses noisy data as much as possible while preserving information about what the data represent. Tishby and Shwartz-Ziv’s new experiments with deep neural networks reveal how the bottleneck procedure actually plays out. In one case, the researchers used small networks that could be trained to label input data with a 1 or 0 (think “dog” or “no dog”) and gave their 282 neural connections random initial strengths. They then tracked what happened as the networks engaged in deep learning with 3,000 sample input data sets.

The basic algorithm used in the majority of deep-learning procedures to tweak neural connections in response to data is called “stochastic gradient descent”: Each time the training data are fed into the network, a cascade of firing activity sweeps upward through the layers of artificial neurons. When the signal reaches the top layer, the final firing pattern can be compared to the correct

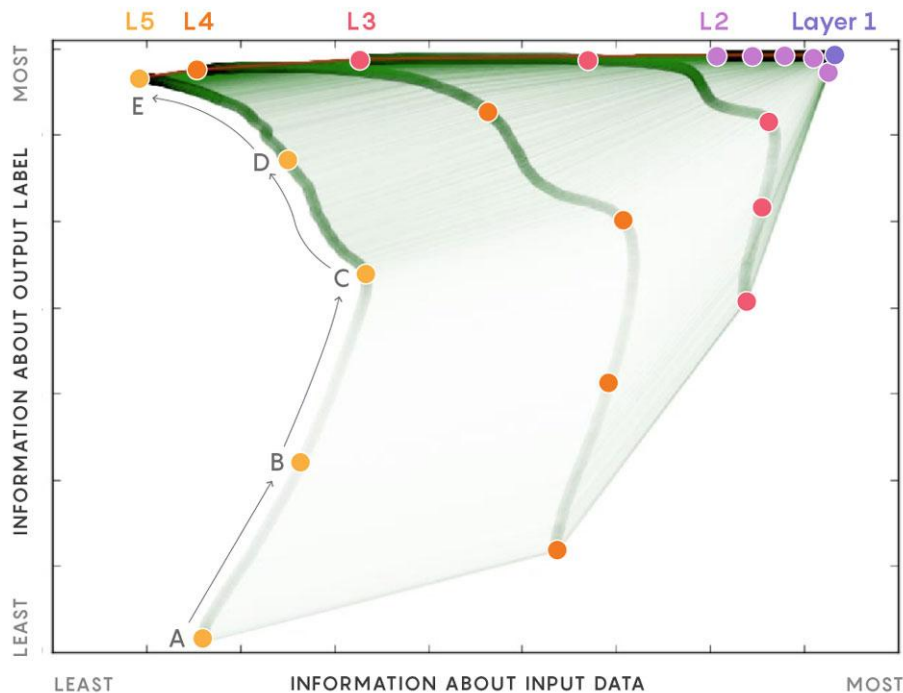
label for the image — 1 or 0, “dog” or “no dog.” Any differences between this firing pattern and the correct pattern are “back-propagated” down the layers, meaning that, like a teacher correcting an exam, the algorithm strengthens or weakens each connection to make the network layer better at producing the correct output signal. Over the course of training, common patterns in the training data become reflected in the strengths of the connections, and the network becomes expert at correctly labeling the data, such as by recognizing a dog, a word, or a 1.

In their experiments, Tishby and Shwartz-Ziv tracked how much information each layer of a deep neural network retained about the input data and how much information each one retained about the output label. The scientists found that, layer by layer, the networks converged to the information bottleneck theoretical bound: a theoretical limit derived in Tishby, Pereira and Bialek’s original paper that represents the absolute best the system can do at extracting relevant information. At the bound, the network has compressed the input as much as possible without sacrificing the ability to accurately predict its label.

Tishby and Shwartz-Ziv also made the intriguing discovery that deep learning proceeds in two phases: a short “fitting” phase, during which the network learns to label its training data, and a much longer “compression” phase, during which it becomes good at generalization, as measured by its performance at labeling new test data.

Inside Deep Learning

New experiments reveal how deep neural networks evolve as they learn.



- A INITIAL STATE:** Neurons in Layer 1 encode everything about the input data, including all information about its label. Neurons in the highest layers are in a nearly random state bearing little to no relationship to the data or its label.
- B FITTING PHASE:** As deep learning begins, neurons in higher layers gain information about the input and get better at fitting labels to it.
- C PHASE CHANGE:** The layers suddenly shift gears and start to “forget” information about the input.
- D COMPRESSION PHASE:** Higher layers compress their representation of the input data, keeping what is most relevant to the output label. They get better at predicting the label.
- E FINAL STATE:** The last layer achieves an optimal balance of accuracy and compression, retaining only what is needed to predict the label.

As a deep neural network tweaks its connections by stochastic gradient descent, at first the number of bits it stores about the input data stays roughly constant or increases slightly, as connections adjust to encode patterns in the input and the network gets good at fitting labels to it. Some experts have compared this phase to memorization.

Then learning switches to the compression phase. The network starts to shed information about the input data, keeping track of only the strongest features — those correlations that are most relevant to the output label. This happens because, in each iteration of stochastic gradient descent, more or less accidental correlations in the training data tell the network to do different things, dialing the strengths of its neural connections up and down in a [random walk](#). This randomization is effectively the same as compressing the system's representation of the input data. As an example, some photos of dogs might have houses in the background, while others don't. As a network cycles through these training photos, it might “forget” the correlation between houses and dogs in some photos as other photos counteract it. It's this forgetting of specifics, Tishby and Shwartz-Ziv argue, that enables the system to form general concepts. Indeed, their experiments revealed that deep neural networks ramp up their generalization performance during the compression phase, becoming better at labeling test data. (A deep neural network trained to recognize dogs in photos might be tested on new photos that may or may not include dogs, for instance.)

It remains to be seen whether the information bottleneck governs all deep-learning regimes, or whether there are other routes to generalization besides compression. Some AI experts see Tishby's idea as one of many important theoretical insights about deep learning to have emerged recently. [Andrew Saxe](#) ([opens a new tab](#)), an AI researcher and theoretical neuroscientist at Harvard University, noted that certain very large deep neural networks don't seem to need a drawn-out compression phase in order to generalize well. Instead, researchers program in something called early stopping, which cuts training short to prevent the network from encoding too many correlations in the first place.

Tishby argues that the network models analyzed by Saxe and his colleagues differ from standard deep neural network architectures, but that nonetheless, the information bottleneck theoretical bound defines these networks' generalization performance better than other methods. Questions about whether the bottleneck holds up for larger neural networks are partly addressed by Tishby and Shwartz-Ziv's most recent experiments, not included in their preliminary paper, in which they train much larger, 330,000-

connection—deep neural networks to recognize handwritten digits in the 60,000-image [Modified National Institute of Standards and Technology database](#)[\(opens a new tab\)](#), a well-known benchmark for gauging the performance of deep-learning algorithms. The scientists saw the same convergence of the networks to the information bottleneck theoretical bound; they also observed the two distinct phases of deep learning, separated by an even sharper transition than in the smaller networks. “I’m completely convinced now that this is a general phenomenon,” Tishby said.

Humans and Machines

The mystery of how brains sift signals from our senses and elevate them to the level of our conscious awareness drove much of the early interest in deep neural networks among AI pioneers, who hoped to reverse-engineer the brain’s learning rules. AI practitioners have since largely abandoned that path in the mad dash for technological progress, instead slapping on bells and whistles that boost performance with little regard for biological plausibility. Still, as their thinking machines achieve ever greater feats — even stoking fears that [AI could someday pose an existential threat](#) — many researchers hope these explorations will uncover general insights about learning and intelligence.

The most important part of learning is actually forgetting.

Naftali Tishby

[Brenden Lake](#)[\(opens a new tab\)](#), an assistant professor of psychology and data science at New York University who studies similarities and differences in how humans and machines learn, said that Tishby’s findings represent “an important step towards opening the black box of neural networks,” but he stressed that the brain represents a much bigger, blacker black box. Our adult brains, which boast several hundred trillion connections between 86 billion neurons, in all likelihood employ a bag of tricks to enhance generalization, going beyond the basic image- and sound-recognition learning procedures that occur during infancy and that may in many ways resemble deep learning. For instance, Lake said the fitting and compression phases that Tishby identified don’t seem to have analogues in the way children learn handwritten characters, which he studies. Children don’t need to see thousands of examples of a character and compress their mental representation over an extended period of time before they’re able to recognize other instances of that letter

and write it themselves. In fact, they can learn from a single example. [Lake and his colleagues' models](#)[\(opens a new tab\)](#) suggest the brain may deconstruct the new letter into a series of strokes — previously existing mental constructs — allowing the conception of the letter to be tacked onto an edifice of prior knowledge. “Rather than thinking of an image of a letter as a pattern of pixels and learning the concept as mapping those features” as in standard machine-learning algorithms, Lake explained, “instead I aim to build a simple causal model of the letter,” a shorter path to generalization.

Such brainy ideas might hold lessons for the AI community, furthering the back-and-forth between the two fields. Tishby believes his information bottleneck theory will ultimately prove useful in both disciplines, even if it takes a more general form in human learning than in AI. One immediate insight that can be gleaned from the theory is a better understanding of which kinds of problems can be solved by real and artificial neural networks. “It gives a complete characterization of the problems that can be learned,” Tishby said. These are “problems where I can wipe out noise in the input without hurting my ability to classify. This is natural vision problems, speech recognition. These are also precisely the problems our brain can cope with.”

Why Artificial Intelligence Like AlphaZero Has Trouble With the Real World

The latest artificial intelligence systems start from zero knowledge of a game and grow to world-beating in a matter of hours. But researchers are struggling to apply these systems beyond the arcade.



Introduction

Update December 6, 2018: This article was originally published in February 2018 based on a paper posted to the scientific preprint site arxiv.org. That paper has now been peer reviewed and published in [Science](#), and we have updated this article to reflect the latest information.

Until very recently, the machines that could trounce champions were at least respectful enough to start by learning from human experience.

To beat Garry Kasparov at chess in 1997, IBM engineers made use of centuries of chess wisdom in their Deep Blue computer. In 2016, Google

DeepMind's [AlphaGo thrashed champion Lee Sedol](#) at the ancient board game Go after poring over millions of positions from tens of thousands of human games.

But now artificial intelligence researchers are rethinking the way their bots incorporate the totality of human knowledge. The current trend is: Don't bother.

In October 2017, the DeepMind team [published](#)[\(opens a new tab\)](#) details of a new Go-playing system, AlphaGo Zero, that studied no human games at all. Instead, it started with the game's rules and played against itself. The first moves it made were completely random. After each game, it folded in new knowledge of what led to a win and what didn't. At the end of these scrimmages, AlphaGo Zero went head to head with the already superhuman version of AlphaGo that had beaten Lee Sedol. It won 100 games to zero.



Lee Sedol, 18-time Go world champion, during his match against AlphaGo in 2016.

The team has now created another master gamer in the AlphaGo family, this one called simply AlphaZero. In a [paper published in Science](#)[\(opens a new tab\)](#), DeepMind researchers revealed that after starting again from scratch, the

trained-up AlphaZero outperformed AlphaGo Zero — in other words, it beat the bot that beat the bot that beat the best Go players in the world. (The paper was first [posted to the scientific preprint site arxiv.org](#)[\(opens a new tab\)](#) in December 2017.) And when it was given the rules for chess or the Japanese chess variant shogi, AlphaZero quickly learned to defeat bespoke top-level algorithms for those games, too. Experts marveled at the program's aggressive, unfamiliar style. "I always wondered how it would be if a superior species landed on Earth and showed us how they played chess," Danish grandmaster Peter Heine Nielsen [told](#)[\(opens a new tab\)](#) a BBC interviewer. "Now I know."

The past year also saw otherworldly self-taught bots emerge in settings as diverse as no-limit poker and Dota 2, a hugely popular multiplayer online video game in which fantasy-themed heroes battle for control of an alien world.

Of course, the companies investing money in these and similar systems have grander ambitions than just dominating video-game tournaments. Research teams like DeepMind hope to apply similar methods to real-world problems like building room-temperature superconductors, or understanding the origami needed to fold proteins into potent drug molecules. And of course, many practitioners hope to eventually build up to artificial general intelligence, an ill-defined but captivating goal in which a machine could think like a person, with the versatility to attack many different kinds of problems.

Yet despite the investments being made in these systems, it isn't yet clear how far past the game board the current techniques can go. "I'm not sure the ideas in AlphaZero generalize readily," said [Pedro Domingos](#)[\(opens a new tab\)](#), a computer scientist at the University of Washington. "Games are a very, very unusual thing."

Perfect Goals for an Imperfect World

One characteristic shared by many games, chess and Go included, is that players can see all the pieces on both sides at all times. Each player always has what's termed "perfect information" about the state of the game. However devilishly complex the game gets, all you need to do is think forward from the current situation.

Plenty of real situations aren't like that. Imagine asking a computer to diagnose an illness or conduct a business negotiation. "Most real-world strategic interactions involve hidden information," said [Noam Brown](#)(opens a new tab), a doctoral student in computer science at Carnegie Mellon University. "I feel like that's been neglected by the majority of the AI community."

Poker, which Brown specializes in, offers a different challenge. You can't see your opponent's cards. But here too, machines that learn by playing against themselves are now reaching superhuman levels. In January 2017, a program called Libratus created by Brown and his adviser, [Tuomas Sandholm](#)(opens a new tab), [outplayed](#)(opens a new tab) four professional poker players at heads-up, no-limit Texas Hold' em, finishing \$1.7 million ahead of its competitors at the end of a 20-day competition.

An even more daunting game involving imperfect information is StarCraft II, another multiplayer online video game with a vast following. Players pick a team, build an army and wage war across a sci-fi landscape. But that landscape is shrouded in a fog of war that only lets players see areas where they have soldiers or buildings. Even the decision to scout your opponent is fraught with uncertainty.

This is one game that AI still can't beat. Barriers to success include the sheer number of moves in a game, which often stretches into the thousands, and the speed at which they must be made. Every player — human or machine — has to worry about a vast set of possible futures with every click.

For now, going toe-to-toe with top humans in this arena is beyond the reach of AI. But it's a target. In August 2017, DeepMind [partnered](#)(opens a new tab) with Blizzard Entertainment, the company that made StarCraft II, to release tools that they say will help open up the game to AI researchers.

Despite its challenges, StarCraft II comes down to a simply enunciated goal: Eradicate your enemy. That's something it shares with chess, Go, poker, Dota 2 and just about every other game. In games, you can win.

From an algorithm's perspective, problems need to have an "objective function," a goal to be sought. When AlphaZero played chess, this wasn't so hard. A loss counted as minus one, a draw was zero, and a win was plus one. AlphaZero's objective function was to maximize its score. The objective function of a poker bot is just as simple: Win lots of money.

Real-life situations are not so straightforward. For example, a self-driving car needs a more nuanced objective function, something akin to the kind of careful phrasing you'd use to explain a wish to a genie. For example: Promptly deliver your passenger to the correct location, obeying all laws and appropriately weighing the value of human life in dangerous and uncertain situations. How researchers craft the objective function, Domingos said, "is one of the things that distinguishes a great machine-learning researcher from an average one."

Consider Tay, a Twitter chatbot released by Microsoft on March 23, 2016. Tay's objective was to engage people, and it did. "What unfortunately Tay discovered," Domingos said, "is that the best way to maximize engagement is to spew out racist insults." It was [snatched back offline](#) [\(opens a new tab\)](#) less than a day later.

Your Own Worst Enemy

Some things don't change. The methods used by today's dominant game bots employ strategies devised decades ago. "It's almost a blast from the past, with just more computation being thrown at it," said [David Duvenaud](#) [\(opens a new tab\)](#), a computer scientist at the University of Toronto.

The strategies often rely on reinforcement learning, a hands-off technique. Instead of micromanaging an algorithm with detailed instructions, engineers let the machine explore an environment and learn to meet goals through trial and error. Before the release of AlphaGo and its progeny, the DeepMind team achieved its first big, headline-grabbing result in 2013, when they used reinforcement learning to make a bot that [learned to play](#) [\(opens a new tab\)](#) seven Atari 2600 games, three of them at an expert level.

That progress has continued. On February 5, DeepMind released [IMPALA](#) [\(opens a new tab\)](#), an AI system that can learn 57 Atari 2600 games, plus 30 more levels built by DeepMind in three dimensions. In these, the player roams through different environments, accomplishing goals like unlocking doors or harvesting mushrooms. IMPALA seems to transfer knowledge between tasks, meaning time spent playing one game also helps it improve at others.

But within the larger category of reinforcement learning, board games and multiplayer games allow for an even more specific approach. Here, exploration can take the form of self-play, where an algorithm gains strategic supremacy by repeatedly wrestling with its own close copy.

This idea dates back decades. In the 1950s, IBM engineer Arthur Samuel [created](#)(opens a new tab) a checkers-playing program that learned in part by matching an alpha side against a beta side. And in the 1990s, Gerald Tesauro, also from IBM, built a backgammon program that pitted the algorithm against itself. The program reached human expert levels, devising unorthodox but effective strategies along the way.

In game after game, an algorithm in a self-play system faces an equally matched foe. This means that changes in strategy lead to different outcomes, giving the algorithm immediate feedback. “Anytime you learn something, anytime you discover a small thing, your opponent immediately uses it against you,” said [Ilya Sutskever](#)(opens a new tab), the research director at OpenAI, a nonprofit he co-founded with Elon Musk that is devoted to developing and sharing AI technology and shepherding it toward safe applications. In August 2017, the organization [released a Dota 2 bot](#)(opens a new tab) controlling the character Shadow Fiend — a sort of demon-necromancer — that beat the world’s best players in one-on-one battles. Another OpenAI project pits simulated humans against one another in a sumo match, where they end up [teaching themselves](#)(opens a new tab) how to tackle and feint. During self-play, “you can never rest, you must always improve,” Sutskever said.

In Dota 2, an online game, a bot designed by OpenAI has taught itself a number of complex strategies.

[OpenAI](#)(opens a new tab)

But the old idea of self-play is just one ingredient in today’s dominant bots, which also need a way to translate their play experiences into deeper understanding. Chess, Go and video games like Dota 2 have many more permutations than there are atoms in the universe. Even over the course of many lifetimes spent battling its own shadow across echoless virtual arenas, a machine can’t encounter every scenario, write it down in a look-up table, and consult that table when it sees the same situation again.

To stay afloat in this sea of possibilities, “you need to generalize, capture the essence,” said [Pieter Abbeel](#)(opens a new tab), a computer scientist at the University of California, Berkeley. IBM’s Deep Blue did this with its built-in chess formula. Armed with the ability to gauge the strength of board positions it hadn’t seen before, it could adopt moves and strategies that would increase its chances of winning. In recent years, though, a new technique has made it possible to skip the formula altogether. “Now, all of a sudden, the ‘deep net’ just captures all of that,” Abbeel said.

Deep neural networks, which have soared in popularity in the last few years, are [built out of layers of artificial “neurons”](#) that stack like pancakes. When neurons in one layer fire, they send signals to the next layer up, which sends them to the next layer, and so on.

By tweaking how the layers connect, these networks become fantastic at morphing inputs into a related output, even if the connection seems abstract. Give them a phrase in English, and they could train themselves to translate it into Turkish. Give them pictures from an animal shelter and they can identify which ones contain cats. Or show them a game board, and they can grok what their probability of winning is. Typically, though, you need to first give these networks reams of labeled examples to practice on.

That’s why self-play and deep neural networks fit together so well. Self-play churns out troves of games, giving deep neural networks a theoretically unlimited supply of the data they need to teach themselves. In turn, the deep neural networks offer a way to internalize the experiences and patterns encountered in self-play.

But there’s a catch. For self-play systems to produce helpful data, they need a realistic place to play in.

“All these games, all of these results, have been in settings where you can perfectly simulate the world,” said [Chelsea Finn](#)[\(opens a new tab\)](#), a Berkeley doctoral student who uses AI to control robot arms and interpret data from sensors. Other domains are not so easy to mock up.

Self-driving cars, for example, have a hard time dealing with bad weather, or cyclists. Or they might not capture bizarre possibilities that turn up in real data, like a bird that happens to fly directly toward the car’s camera. For robot arms, Finn said, initial simulations provide basic physics, allowing the arm to at least learn how to learn. But they fail to capture the details involved in touching surfaces, which means that tasks like screwing on a bottle cap — or conducting an intricate surgical procedure — require real-world experience, too.

For problems that are hard to simulate, then, self-play is not so useful. “There is a huge difference between a true perfect model of the environment and a learned estimated one, especially when that reality is complex,” wrote [Yoshua](#)

[Bengio](#)(opens a new tab), a pioneer of deep learning at the University of Montreal, in an email. But that still leaves AI researchers with ways to move forward.

Life After Games

It's hard to pinpoint the dawn of AI gaming supremacy. You could choose Kasparov's loss in chess, or Lee Sedol's defeat at the virtual hands of AlphaGo. Another popular option would be when legendary [Jeopardy!](#) champion Ken Jennings lost to IBM's Watson in 2011. Watson could parse the game's clues and handle wordplay. The two-day match wasn't close. "I for one welcome our new computer overlords," Jennings wrote under his final answer.

Watson seemed to be endowed with the kind of clerical skills humans use on a host of real-world problems. It could take a prompt in English, rummage through relevant documents at lightning speed, come up with the relevant snippets of information, and settle on a single best answer. But seven years later, the real world continues to present stubborn challenges for AI. A September [report by the health publication Stat](#)(opens a new tab) found that researching and designing personalized cancer treatments, as Watson's descendant Watson for Oncology aims to do, is proving difficult.

"The questions in [Jeopardy!](#) are easier in the sense that they don't need much common sense," wrote Bengio, who has collaborated with the Watson team, when asked to compare the two cases from the AI perspective. "Understanding a medical article is much harder. Again, much basic research is needed."

As special as games are, there are still a few real-world problems they resemble. Researchers from DeepMind declined to be interviewed for this article, citing the fact that their AlphaZero work is currently under peer review. But the team has suggested that its techniques may soon help biomedical researchers, who would like to understand protein folding.

To do this, they need to figure out how the various amino acids that make up a protein [kink and fold](#)(opens a new tab) into a little three-dimensional machine with a function that depends on its shape. That's tricky in the same ways chess is tricky: Chemists know the rules roughly well enough to calculate specific scenarios, but there are still so many possible configurations, it's a hopeless task to search through them all. But what if protein folding could be configured

as a game? In fact, it already has been. Since 2008, hundreds of thousands of human players have attempted [Foldit](#)(opens a new tab), an online game where users are scored on the stability and feasibility of the protein structures they fold. A machine could train itself in a similar manner, perhaps by trying to beat its previous best score with general reinforcement learning.

Reinforcement learning and self-play might also help train dialogue systems, Sutskever suggests. That would give robots meant to speak to humans a chance to train by talking to themselves. And considering that specialized AI hardware is becoming faster and more available, engineers will have an incentive to pose more and more problems in the form of games. “I think that in the future, self-play and other ways of consuming a very large amount of computing power will become more and more important,” Sutskever said.

But if the ultimate goal is for machines to do as much as humans can, even a self-taught, generalist board-game champ like AlphaZero may have a ways to go. “You have to see, to my mind at least, what’s really a huge gulf between the real activities of thinking, creative exploration of ideas, and what we currently see in AI,” said [Josh Tenenbaum](#)(opens a new tab), a cognitive scientist at the Massachusetts Institute of Technology. “That kind of intelligence is there, but it’s mostly going on in the minds of the great AI researchers.”

Many other researchers, conscious of the hype that surrounds their field, offer their own qualifiers. “I would be careful not to overestimate the significance of playing these games, for AI or jobs in general. Humans are not very good at games,” said François Chollet, a deep-learning researcher at Google.

“But keep in mind that very simple, specialized tools can actually achieve a lot,” he said.

Clarification February 22, 2018: An earlier version of this article implied that chess strategy was programmed into the Deep Blue computer. In fact, engineers programmed in a framework for chess strategy, and the machine analyzed many human games to arrive at its own particular strategy. The article has been changed to avoid the unwanted implication.

Artificial Intelligence Learns to Learn Entirely on Its Own

A new version of AlphaGo needed no human instruction to figure out how to clobber the best Go player in the world — itself.

Introduction

Amere 19 months after dethroning the world's top human Go player, the computer program AlphaGo has smashed an even more momentous barrier: It can now achieve unprecedented levels of mastery purely by teaching itself. Starting with zero knowledge of Go strategy and no training by humans, the new iteration of the program, called AlphaGo Zero, needed just three days to invent advanced strategies undiscovered by human players in the multi-millennia history of the game. By freeing artificial intelligence from a dependence on human knowledge, the breakthrough removes a primary limit on how smart machines can become.

Earlier versions of AlphaGo were [taught to play the game using two methods](#). In the first, called supervised learning, researchers fed the program 100,000 top amateur Go games and taught it to imitate what it saw. In the second, called reinforcement learning, they had the program play itself and learn from the results.

AlphaGo Zero skipped the first step. The program began as a blank slate, knowing only the rules of Go, and played games against itself. At first, it placed stones randomly on the board. Over time it got better at evaluating board positions and identifying advantageous moves. It also learned many of the canonical elements of Go strategy and discovered new strategies all its own. “When you learn to imitate humans the best you can do is learn to imitate humans,” said [Satinder Singh](#) ([opens a new tab](#)), a computer scientist at a University of Michigan who was not involved with the research. “In many complex situations there are new insights you’ll never discover.”

After three days of training and 4.9 million training games, the researchers matched AlphaGo Zero against the earlier champion-beating version of the program. AlphaGo Zero won 100 games to zero.

To expert observers, the rout was stunning. Pure reinforcement learning would seem to be no match for the overwhelming number of possibilities in Go, which is vastly more complex than chess: You'd have expected AlphaGo Zero to spend forever searching blindly for a decent strategy. Instead, it rapidly found its way to superhuman abilities.

The efficiency of the learning process owes to a feedback loop. Like its predecessor, AlphaGo Zero determines what move to play through a process called a "tree search." The program starts with the current board and considers the possible moves. It then considers what moves its opponent could play in each of the resulting boards, and then the moves it could play in response and so on, creating a branching tree diagram that simulates different combinations of play resulting in different board setups. DeepMind

AlphaGo Zero can't follow every branch of the tree all the way through, since that would require inordinate computing power. Instead, it selectively prunes branches by deciding which paths seem most promising. It makes that calculation — of which paths to prune — based on what it has learned in earlier play about the moves and overall board setups that lead to wins.

Earlier versions of AlphaGo did all this, too. What's novel about AlphaGo Zero is that instead of just running the tree search and making a move, it remembers the outcome of the tree search — and eventually of the game. It then uses that information to update its estimates of promising moves and the probability of winning from different positions. As a result, the next time it runs the tree search it can use its improved estimates, trained with the results of previous tree searches, to generate even better estimates of the best possible move.

The computational strategy that underlies AlphaGo Zero is effective primarily in situations in which you have an extremely large number of possibilities and want to find the optimal one. In [the Nature paper describing the research](#) (opens a new tab), the authors of AlphaGo Zero suggest that their system could be useful in materials exploration — where you want to identify atomic combinations that yield materials with different properties — and protein folding, where you want to understand how a protein's precise three-dimensional structure determines its function.

As for Go, the effects of AlphaGo Zero are likely to be seismic. To date, gaming companies have failed in their efforts to develop world-class Go software. AlphaGo Zero is likely to change that. [Andrew Jackson](#) (opens a new tab), executive vice president of the American Go Association, thinks it won't be

long before Go apps appear on the market. This will change the way human Go players train. It will also make cheating easier.

As for AlphaGo, the future is wide open. Go is sufficiently complex that there's no telling how good a self-starting computer program can get; and AlphaGo now has a learning method to match the expansiveness of the game it was bred to protect.

[artificial intelligence](#) [biology](#) [cyber](#) [robotics](#) [science](#) [machine learning](#) [neural](#)

Mapping the Brain to Build Better Machines

A race to decipher the brain's algorithms could revolutionize machine learning.



Take a three year-old to the zoo, and she intuitively knows that the long-necked creature nibbling leaves is the same thing as the giraffe in her picture book. That superficially easy feat is in reality quite sophisticated. The cartoon drawing is a frozen silhouette of simple lines, while the living animal is awash in color, texture, movement and light. It can contort into different shapes and looks different from every angle.

Humans excel at this kind of task. We can effortlessly grasp the most important features of an object from just a few examples and apply those features to the unfamiliar. Computers, on the other hand, typically need to sort through a whole database of giraffes, shown in many settings and from different perspectives, to learn to accurately recognize the animal.

Visual identification is one of many arenas where humans beat computers. We're also better at finding relevant information in a flood of data; at solving unstructured problems; and at learning without supervision, as a baby learns about gravity when she plays with blocks. "Humans are much, much better generalists," said [Tai Sing Lee\(opens a new tab\)](#), a computer scientist and neuroscientist at Carnegie Mellon University in Pittsburgh. "We are still more flexible in thinking and can anticipate, imagine and create future events."

An ambitious new program, funded by the federal government's intelligence arm, aims to bring artificial intelligence more in line with our own mental powers. Three teams composed of neuroscientists and computer scientists will attempt to figure out how the brain performs these feats of visual identification, then make machines that do the same. "Today's machine learning fails where humans excel," said [Jacob Vogelstein\(opens a new tab\)](#), who heads the program at the Intelligence Advanced Research Projects Activity (IARPA). "We want to revolutionize machine learning by reverse engineering the algorithms and computations of the brain."

Time is short. Each team is now modeling a chunk of cortex in unprecedented detail. In conjunction, the teams are developing algorithms based in part on what they learn. By next summer, each of those algorithms will be given an example of a foreign item and then required to pick out instances of it from among thousands of images in an unlabeled database. "It is a very aggressive time-frame," said [Christof Koch\(opens a new tab\)](#), president and chief scientific officer of the Allen Institute for Brain Science in Seattle, which is working with one of the teams.

Koch and his colleagues are now creating a complete wiring diagram of a small cube of brain — a million cubic microns, totaling one five-hundredth the volume of a poppy seed. That's orders of magnitude larger than the most-extensive complete wiring map to date, which was published [last June\(opens a new tab\)](#) and took roughly six years to complete.

By the end of the five-year IARPA project, dubbed [Machine Intelligence from Cortical Networks\(opens a new tab\)](#) (Microns), researchers aim to map a cubic millimeter of cortex. That tiny portion houses about 100,000 neurons, 3 to 15 million neuronal connections, or synapses, and enough neural wiring to span the width of Manhattan, were it all untangled and laid end-to-end.

No one has yet attempted to reconstruct a piece of brain at this scale. But smaller-scale efforts have shown that these maps can provide insight into the inner workings of the cortex. In a paper published in the journal [Nature](#) in March, Wei-Chung Allen Lee — a neuroscientist at Harvard University who is working with Koch's team — and his collaborators [mapped out\(opens a new tab\)](#) a wiring diagram of 50 neurons and more than 1,000 of their partners. By pairing this map with information about each neuron's job in the brain — some respond to a visual input of vertical bars, for example — they derived a simple rule for how neurons in this part of the cortex are anatomically connected. They found that neurons with similar functions are more likely to both connect to and make larger connections with each other than they are with other neuron types.

While the implicit goal of the Microns project is technological — IARPA funds research that could eventually lead to data-analysis tools for the intelligence community, among other things

— new and profound insights into the brain will have to come first. [Andreas Tolias](#)(opens a new tab), a neuroscientist at Baylor College of Medicine who is co-leading Koch’s team, likens our current knowledge of the cortex to a blurry photograph. He hopes that the unprecedented scale of the Microns project will help sharpen that view, exposing more sophisticated rules that govern our neural circuits. Without knowing all the component parts, he said, “maybe we’re missing the beauty of the structure.”

The Brain’s Processing Units

The convoluted folds covering the brain’s surface form the cerebral cortex, a pizza-sized sheet of tissue that’s scrunched to fit into our skulls. It is in many ways the brain’s microprocessor. The sheet, roughly three millimeters thick, is made up of a series of repeating modules, or microcircuits, similar to the array of logic gates in a computer chip. Each module consists of approximately 100,000 neurons arranged in a complex network of interconnected cells. Evidence suggests that the basic structure of these modules is roughly the same throughout the cortex. However, modules in different brain regions are specialized for specific purposes such as vision, movement or hearing.

Scientists have only a rough sense of what these modules look like and how they act. They’ve largely been limited to studying the brain at smaller scales: tens or hundreds of neurons. New technologies designed to trace the shape, activity and connectivity of thousands of neurons are finally allowing researchers to analyze how cells within a module interact with each other; how activity in one part of the system might spark or dampen activity in another part. “For the first time in history, we have the ability to interrogate the modules instead of just guessing at the contents,” Vogelstein said. “Different teams have different guesses for what’s inside.”

The researchers will focus on a part of the cortex that processes vision, a sensory system that neuroscientists have explored intensively and that computer scientists have long striven to emulate. “Vision seems easy — just open your eyes — but it’s hard to teach computers to do the same thing,” said [David Cox](#)(opens a new tab), a neuroscientist at Harvard who leads one of the IARPA teams.

Each team is starting with the same basic idea for how vision works, a decades-old theory known as analysis-by-synthesis. According to this idea, the brain makes predictions about what will happen in the immediate future and then reconciles those predictions with what it sees. The power of this approach lies in its efficiency — it requires less computation than continuously recreating every moment in time.

The brain might execute analysis-by-synthesis any number of different ways, so each team is exploring a different possibility. Cox’s team views the brain as a sort of physics engine, with existing physics models that it uses to simulate what the world should look like. Tai Sing Lee’s team, co-led by [George Church](#)(opens a new tab), theorizes that the brain has built a library of

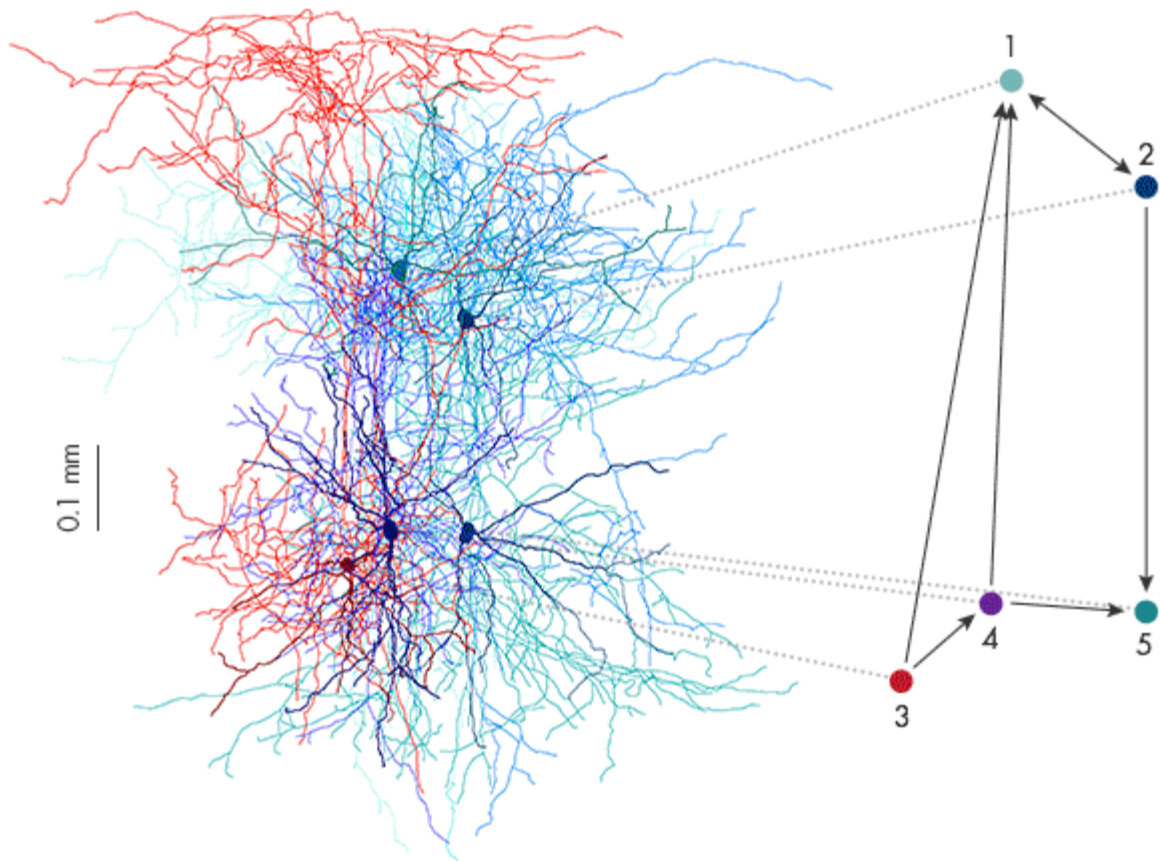
parts — bits and pieces of objects and people — and learns rules for how to put those parts together. Leaves, for example, tend to appear on branches. Tolias’s group is working on a more data-driven approach, where the brain creates statistical expectations of the world in which it lives. His team will test various hypotheses for how different parts of the circuit learn to communicate.

All three teams will monitor neuronal activity from tens of thousands of neurons in a target cube of brain. Then they’ll use different methods to create a wiring diagram of those cells. Cox’s team, for example, will slice brain tissue into layers thinner than a human hair and analyze each slice with electron microscopy. The team will then computationally stitch together each cross section to create a densely packed three-dimensional map that charts millions of neural wires on their intricate path through the cortex.

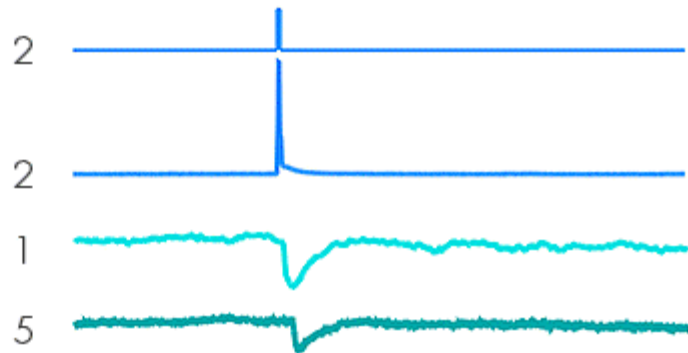
With a map and activity pattern in hand, each team will attempt to tease out some basic rules governing the circuit. They’ll then program those rules into a simulation and measure how well the simulation matches a real brain.

Tolias and collaborators already have a taste of what this type of approach can accomplish. In a [paper published \(opens a new tab\)](#) in *Science* in November, they mapped the connections between 11,000 neuronal pairs, uncovering five new types of neurons in the process. “We still don’t have a complete listing of the parts that make up cortex, what the individual cells look like, how they are connected,” said Koch. “That’s what [Tolias] has started to do.”

MAPPING THE CORTEX



ELECTRICAL
ACTIVITY



Andreas Tolias and collaborators mapped out the connections among pairs of neurons and recorded their electrical activity. The complex anatomy of five neurons (top-left) can be boiled down to a simple circuit diagram (top-right). Injecting electrical current into neuron 2 makes the neuron fire, triggering electrical changes in the two cells downstream, neurons 1 and 5 (bottom).

Olena Shmahalo/Quanta Magazine; Andreas Tolias

Among these thousands of neuronal connections, Tolias’s team uncovered three general rules that govern how the cells are connected: Some talk mainly to neurons of their own kind; others avoid their own kind, communicating mostly with other varieties; and a third group talks only to a few other neurons. (Tolias’s team defined their cells based on neural anatomy rather than function, which Wei Lee’s team did in their study.) Using just these three wiring rules, the researchers could simulate the circuit fairly accurately. “Now the challenge is to figure out what those wiring rules mean algorithmically,” Tolias said. “What kinds of calculations do they do?”

Neural Nets Based on Real Neurons

Brain-like artificial intelligence isn’t a new idea. So-called neural networks, which mimic the basic structure of the brain, were extremely popular in the 1980s. But at the time, the field lacked the computing power and training data that the algorithms needed to become really effective. All of the Internet’s millions of labeled cat pictures weren’t yet available, after all. And although neural networks have enjoyed a major renaissance — the voice- and face-recognition programs that have rapidly become part of our daily lives are based on neural network algorithms, as is [AlphaGo](#), the computer that recently defeated the world’s top Go player — the rules that artificial neural networks use to alter their connections are almost certainly different than the ones employed by the brain.

Contemporary neural networks “are based on what we knew about the brain in the 1960s,” said [Terry Sejnowski](#)[\(opens a new tab\)](#), a computational neuroscientist at the Salk Institute in San Diego who developed early neural network algorithms with [Geoffrey Hinton](#)[\(opens a new tab\)](#), a computer scientist at the University of Toronto. “Our knowledge of how the brain is organized is exploding.”

For example, today’s neural networks are comprised of a feed-forward architecture, where information flows from input to output through a series of layers. Each layer is trained to recognize certain features, such as an eye or a whisker. That analysis is then fed forward, with each successive layer performing increasingly complex computations on the data. In this way, the program eventually recognizes a series of colored pixels as a cat.

But this feed-forward structure leaves out a vital component of the biological system: feedback, both within individual layers and from higher-order layers to lower-order ones. In the real brain, neurons in one layer of the cortex are connected to their neighbors, as well as to neurons in the layers above and below them, creating an intricate network of loops. “Feedback connections are an incredibly important part of cortical networks,” Sejnowski said. “There are as many feedback as feed-forward connections.”

Neuroscientists don’t yet precisely understand what these feedback loops are doing, though they know they are important for our ability to direct our attention. They help us listen to a voice on the phone while tuning out distracting city sounds, for example. Part of the appeal of the

analysis-by-synthesis theory is that it provides a reason for all those recurrent connections. They help the brain compare its predictions with reality.

Microns researchers aim to decipher the rules governing feedback loops — such as which cells these loops connect, what triggers their activity, and how that activity effects the circuit’s output — then translate those rules into an algorithm. “What is lacking in a machine right now is imagination and introspection. I believe the feedback circuitry allows us to imagine and introspect at many different levels,” Tai Sing Lee said.

Perhaps feedback circuitry will one day endow machines with traits we think of as uniquely human. “If you could implement [feedback circuitry] in a deep network, you could go from a network that has kind of a knee-jerk reaction — give input and get output — to one that’s more reflective, that can start thinking about inputs and testing hypotheses,” said Sejnowski, who serves as an advisor to President Obama’s \$100 million [BRAIN Initiative](#)[\(opens a new tab\)](#), of which the Microns project is a part.

Clues to Consciousness

Like all IARPA programs, the Microns project is high risk. The technologies that researchers need for large-scale mapping of neuronal activity and wiring exist, but no one has applied them at this scale before. One challenge will be dealing with the [enormous amounts of data](#) the research produces — 1 to 2 petabytes of data per millimeter cube of brain. The teams will likely need to develop new machine-learning tools to analyze all that data, a rather ironic feedback loop of its own.

It’s also unclear whether the lessons learned from a small chunk of brain will prove illustrative of the brain’s larger talents. “The brain isn’t just a piece of cortex,” Sejnowski said. “The brain is hundreds of systems specialized for different functions.” The cortex itself is made up of repeating units that look roughly the same. But other parts of the brain might act quite differently. The reinforcement learning employed in the [AlphaGo](#) algorithm, for example, is related to processes that take place in the basal ganglia, part of the brain involved in addiction. “If you want AI that goes beyond simple pattern recognition, you’re going to need a lot of different parts,” Sejnowski said.

Should the project succeed, however, it will do more than analyze intelligence data. A successful algorithm will reveal important truths about how the brain makes sense of the world. In particular, it will help confirm that the brain does indeed operate via analysis-by-synthesis — that it compares its own predictions about the world with the incoming data washing through our senses. It will reveal that a key ingredient in the recipe for consciousness is an ever-shifting mixture of imagination plus perception. “It is imagination that allows us to predict future events and use that to guide our actions,” Tai Sing Lee said. By building machines that think, these researchers hope to reveal the secrets of thought itself.

