The Case That A.I. Is Thinking

ChatGPT does not have an inner life. Yet it seems to know what it's talking about.

By James Somers November 3, 2025

How convincing does the illusion of understanding have to be before you stop calling it an illusion?Animations by Zach Lieberman

Dario Amodei, the C.E.O. of the artificial-intelligence company Anthropic, has been predicting that an A.I. "smarter than a Nobel Prize winner" in such fields as biology, math, engineering, and writing might come online by 2027. He envisions millions of copies of a model whirring away, each conducting its own research: a "country of geniuses in a datacenter." In June, Sam Altman, of OpenAI, wrote that the industry was on the cusp of building "digital superintelligence." "The 2030s are likely going to be wildly different from any time that has come before," he asserted. Meanwhile, the A.I. tools that most people currently interact with on a day-to-day basis are reminiscent of Clippy, the onetime Microsoft Office "assistant" that was actually more of a gadfly. A Zoom A.I. tool suggests that you ask it "What are some meeting icebreakers?" or instruct it to "Write a short message to share gratitude." Siri is good at setting

reminders but not much else. A friend of mine saw a button in Gmail that said "Thank and tell anecdote." When he clicked it, Google's A.I. invented a funny story about a trip to Turkey that he never took.

The rushed and uneven rollout of A.I. has created a fog in which it is tempting to conclude that there is nothing to see here—that it's all hype. There is, to be sure, plenty of hype: Amodei's timeline is science-fictional. (A.I. models aren't improving that fast.) But it is another kind of wishful thinking to suppose that large language models are just shuffling words around. I used to be sympathetic to that view. I sought comfort in the idea that A.I. had little to do with real intelligence or understanding. I even celebrated its shortcomings—rooting for the home team. Then I began using A.I. in my work as a programmer, fearing that if I didn't I would fall behind. (My employer, a trading firm, has several investments in and partnerships with A.I. companies, including Anthropic.) Writing code is, by many accounts, the thing that A.I. is best at; code has more structure than prose does, and it's often possible to automatically validate that a given program works. My conversion was swift. At first, I consulted A.I. models in lieu of looking something up. Then I gave them small, self-contained problems. Eventually, I gave them real work—the kind I'd trained my whole career to do. I saw these models digest, in seconds, the intricate details of thousands of lines of code. They could spot subtle bugs and

orchestrate complex new features. Finally, I was transferred to a fast-growing team that aims to make better use of A.I. tools, and to create our own.

The science-fiction author William Gibson is said to have observed that the future is already here, just not evenly distributed—which might explain why A.I. seems to have minted two cultures, one dismissive and the other enthralled. In our daily lives, A.I. "agents" that can book vacations or file taxes are a flop, but I have colleagues who compose much of their code using A.I. and sometimes run multiple coding agents at a time. Models sometimes make amateur mistakes or get caught in inane loops, but, as I've learned to use them effectively, they have allowed me to accomplish in an evening what used to take a month. Not too long ago, I made two iOS apps without knowing how to make an iOS app.

"O.K., we're good on bread crumbs. Now we're looking for a pound of ground beef and a pound of veal."

Cartoon by Olivia Noble

I once had a boss who said that a job interview should probe for strengths, not for the absence of weaknesses. Large language models have many weaknesses: they famously hallucinate reasonable-sounding falsehoods; they can be servile even when you're wrong; they are fooled by simple puzzles. But I remember a time when the obvious strengths of today's A.I. models—fluency, fluidity, an ability to "get"

what someone is talking about—were considered holy grails. When you experience these strengths firsthand, you wonder: How convincing does the illusion of understanding have to be before you stop calling it an illusion?

On a brutally hot day this summer, my friend Max met up with his family at a playground. For some reason, a sprinkler for kids was switched off, and Max's wife had promised everyone that her husband would fix it. Confronted by red-faced six- and seven-year-olds, Max entered a utility shed hoping to find a big, fat "On" switch. Instead, he found a maze of ancient pipes and valves. He was about to give up when, on a whim, he pulled out his phone and fed a photo into ChatGPT-4o, along with a description of his problem. The A.I. thought for a second, or maybe didn't think, but all the same it said that he was looking at a backflow-preventer system typical of irrigation setups. Did he see that yellow ball valve toward the bottom? That probably controlled the flow. Max went for it, and cheers rang out across the playground as the water turned on.

Get our Classics newsletter to discover timeless gems from *The New Yorker* archive.

Was ChatGPT mindlessly stringing words together, or did it understand the problem? The answer could teach us something important about understanding itself. "Neuroscientists have to confront this humbling truth," Doris

Tsao, a neuroscience professor at the University of California, Berkeley, told me. “The advances in machine learning have taught us more about the essence of intelligence than anything that neuroscience has discovered in the past hundred years.” Tsao is best known for decoding how macaque monkeys perceive faces. Her team learned to predict which neurons would fire when a monkey saw a specific face; even more strikingly, given a pattern of neurons firing, Tsao’s team could render the face. Their work built on research into how faces are represented inside A.I. models. These days, her favorite question to ask people is “What is the deepest insight you have gained from ChatGPT?” “My own answer,” she said, “is that I think it radically demystifies thinking.”

The most basic account of how we got here goes something like this. In the nineteen-eighties, a small team of cognitive psychologists and computer scientists tried to simulate thinking in a machine. Among the more famous of them were David Rumelhart, Geoffrey Hinton, and James McClelland, who went on to form a research group at U.C. San Diego. They saw the brain as a vast network in which neurons fire in patterns, causing other sets of neurons to fire, and so on; this dance of patterns is thinking. The brain learns by changing the strength of the connections between neurons. Crucially, the scientists mimicked this process by creating an artificial neural network, and by applying a simple algorithm

called gradient descent to increase the accuracy of its predictions. (The algorithm could be compared to a hiker navigating from a mountaintop to a valley; a simple strategy for eventually finding one's way is to insure that every step moves downhill.) The use of such algorithms in large networks is known as deep learning.

Other people in A.I. were skeptical that neural networks were sophisticated enough for real-world tasks, but, as the networks got bigger, they began to solve previously unsolvable problems. People would devote entire dissertations to developing techniques for distinguishing handwritten digits or for recognizing faces in images; then a deep-learning algorithm would digest the underlying data, discover the subtleties of the problem, and make those projects seem obsolete. Deep learning soon conquered speech recognition, translation, image captioning, board games, and even the problem of predicting how proteins will fold.

Today's leading A.I. models are trained on a large portion of the internet, using a technique called next-token prediction. A model learns by making guesses about what it will read next, then comparing those guesses to whatever actually appears. Wrong guesses inspire changes in the connection strength between the neurons; this is gradient descent. Eventually, the model becomes so good at predicting text

that it appears to know things and make sense. So that is something to think about. A group of people sought the secret of how the brain works. As their model grew toward a brain-like size, it started doing things that were thought to require brain-like intelligence. Is it possible that they found what they were looking for?

There is understandable resistance to such a simplistic and triumphant account of A.I. The case against it was well argued by Ted Chiang, who wrote an article for this magazine in early 2023 titled "ChatGPT Is a Blurry *JPEG* of the Web." He meant it in a more or less deflationary way: that's *all* ChatGPT is. You feed the whole internet to a program and it regurgitates it back to you imperfectly, like a copy of a copy of a photograph—but with just enough facility to fool you into believing that the program is intelligent. This spring, a similar argument was made in a book, "The AI Con," by Emily M. Bender, a linguist, and Alex Hanna, a sociologist. Bender is perhaps best known for describing L.L.M.s as "stochastic parrots." "Large language models do not, cannot, and will not 'understand' anything at all," the writer Tyler Austin Harper declared in a book review in *The Atlantic*. Models "produce writing not by thinking but by making statistically informed guesses about which lexical item is likely to follow another." Harper buttressed these technical arguments with moral ones. A.I. enriches the powerful, consumes enough energy to accelerate climate change, and

marginalizes workers. He concluded that “the foundation of the AI industry is a scam.”

A leading neuroscientist argues that ChatGPT “radically demystifies thinking.”

But the moral case against A.I. may ultimately be stronger than the technical one. “The ‘stochastic parrot’ thing has to be dead at some point,” Samuel J. Gershman, a Harvard cognitive scientist who is no A.I. hype man, told me. “Only the most hardcore skeptics can deny these systems are doing things many of us didn’t think were going to be achieved.” Jonathan Cohen, a cognitive neuroscientist at Princeton, emphasized the limitations of A.I., but argued that, in some cases, L.L.M.s seem to mirror one of the largest and most important parts of the human brain. “To a first approximation, your neocortex is your deep-learning mechanism,” Cohen said. Humans have a much larger neocortex than other animals, relative to body size, and the species with the largest neocortices—elephants, dolphins, gorillas, chimpanzees, dogs—are among the most intelligent.

In 2003, the machine-learning researcher Eric B. Baum published a book called “What Is Thought?” (I stumbled upon it in my college’s library stacks, drawn by the title.) The gist of Baum’s argument is that understanding is compression, and compression is understanding. In statistics, when you want to make sense of points on a graph, you can use a technique called linear regression to

draw a “line of best fit” through them. If there’s an underlying regularity in the data—maybe you’re plotting shoe size against height—the line of best fit will efficiently express it, predicting where new points could fall. The neocortex can be understood as distilling a sea of raw experience—sounds, sights, and other sensations—into “lines of best fit,” which it can use to make predictions. A baby exploring the world tries to guess how a toy will taste or where food will go when it hits the floor. When a prediction is wrong, the connections between neurons are adjusted. Over time, those connections begin to capture regularities in the data. They form a compressed model of the world.

Artificial neural networks compress experience just like real neural networks do. One of the best open-source A.I. models, DeepSeek, is capable of writing novels, suggesting medical diagnoses, and sounding like a native speaker in dozens of languages. It was trained using next-token prediction on many terabytes of data. But when you download the model it is one six-hundredth of that. A distillation of the internet, compressed to fit on your laptop. Ted Chiang was right to call an early version of ChatGPT a blurry *JPEG* of the web—but, in my view, this is the very reason these models have become increasingly intelligent. Chiang noted in his piece that, to compress a text file filled with millions of examples of arithmetic, you wouldn’t create a zip file. You’d write a calculator program. “The greatest

degree of compression can be achieved by understanding the text," he wrote. Perhaps L.L.M.s are starting to do that.

It can seem unnatural, even repulsive, to imagine that a computer program actually understands, actually *thinks*. We usually conceptualize thinking as something conscious, like a Joycean inner monologue or the flow of sense memories in a Proustian daydream. Or we might mean reasoning: working through a problem step by step. In our conversations about A.I., we often conflate these different kinds of thinking, and it makes our judgments pat. ChatGPT is obviously not thinking, goes one argument, because it is obviously not having a Proustian reverie; ChatGPT clearly is thinking, goes another, because it can work through logic puzzles better than you can.

Something more subtle is going on. I do not believe that ChatGPT has an inner life, and yet it seems to know what it's talking about. Understanding—having a grasp of what's going on—is an underappreciated kind of thinking, because it's mostly unconscious. Douglas Hofstadter, a professor of cognitive science and comparative literature at Indiana University, likes to say that cognition is recognition. Hofstadter became famous for a book about the mind and consciousness called "Gödel, Escher, Bach: An Eternal Golden Braid," which won a Pulitzer Prize in 1980. Hofstadter's theory, developed through decades of

research, is that “seeing as” is the essence of thinking. You see one patch of color as a car and another as a key chain; you recognize the letter “A” no matter what font it is written in or how bad the handwriting might be. Hofstadter argued that the same process underlies more abstract kinds of perception. When a grand master examines a chess board, years of practice are channelled into a way of seeing: white’s bishop is weak; that endgame is probably a draw. You see an eddy in a river as a sign that it’s dangerous to cross. You see a meeting you’re in as an emperor-has-no-clothes situation. My nearly two-year-old son recognizes that late-morning stroller walks might be an opportunity for a croissant and makes demands accordingly. For Hofstadter, that’s intelligence in a nutshell.

Hofstadter was one of the original A.I. deflationists, and my own skepticism was rooted in his. He wrote that most A.I. research had little to do with real thinking, and when I was in college, in the two-thousands, I agreed with him. There were exceptions. He found the U.C.S.D. group interesting. And he admired the work of a lesser-known Finnish American cognitive scientist, Pentti Kanerva, who noticed some unusual properties in the mathematics of high-dimensional spaces. In a high-dimensional space, any two random points may be extremely far apart. But, counterintuitively, each point also has a large cloud of neighbors around it, so you can easily find your way to it if you get “close enough.” That

reminded Kanerva of the way that memory works. In a 1988 book called "Sparse Distributed Memory," Kanerva argued that thoughts, sensations, and recollections could be represented as coördinates in high-dimensional space. The brain seemed like the perfect piece of hardware for storing such things. Every memory has a sort of address, defined by the neurons that are active when you recall it. New experiences cause new sets of neurons to fire, representing new addresses. Two addresses can be different in many ways but similar in others; one perception or memory triggers other memories nearby. The scent of hay recalls a memory of summer camp. The first three notes of Beethoven's Fifth beget the fourth. A chess position that you've never seen reminds you of old games—not all of them, just the ones in the right neighborhood.

"Bye, sweetie—have a day filled with social drama, drastically shifting friendships, and academic milestones, which you'll describe to me later as 'fine.' "

Cartoon by Ali Solomon

Hofstadter realized that Kanerva was describing something like a "seeing as" machine. "Pentti Kanerva's memory model was a revelation for me," he wrote in a foreword to Kanerva's book. "It was the very first piece of research I had ever run across that made me feel I could glimpse the distant goal of understanding how the brain works as a whole." Every kind of thinking—whether Joycean, Proustian, or logical—

depends on the relevant thing coming to mind at the right time. It's how we figure out what situation we're in.

Kanerva's book receded from view, and Hofstadter's own star faded—except when he occasionally poked up his head to criticize a new A.I. system. In 2018, he wrote of Google Translate and similar technologies: "There is still something deeply lacking in the approach, which is conveyed by a single word: *understanding*." But GPT-4, which was released in 2023, produced Hofstadter's conversion moment. "I'm mind-boggled by some of the things that the systems do," he told me recently. "It would have been inconceivable even only ten years ago." The staunchest deflationist could deflate no longer. Here was a program that could translate as well as an expert, make analogies, extemporize, generalize. Who were we to say that it didn't understand? "They do things that are very much like thinking," he said. "You could say they *are* thinking, just in a somewhat alien way."

L.L.M.s appear to have a "seeing as" machine at their core. They represent each word with a series of numbers denoting its coördinates—its vector—in a high-dimensional space. In GPT-4, a word vector has thousands of dimensions, which describe its shades of similarity to and difference from every other word. During training, a large language model tweaks a word's coördinates whenever it makes a prediction error; words that appear in texts together are nudged closer in

space. This produces an incredibly dense representation of usages and meanings, in which analogy becomes a matter of geometry. In a classic example, if you take the word vector for “Paris,” subtract “France,” and then add “Italy,” the nearest other vector will be “Rome.” L.L.M.s can “vectorize” an image by encoding what’s in it, its mood, even the expressions on people’s faces, with enough detail to redraw it in a particular style or to write a paragraph about it. When Max asked ChatGPT to help him out with the sprinkler at the park, the model wasn’t just spewing text. The photograph of the plumbing was compressed, along with Max’s prompt, into a vector that captured its most important features. That vector served as an address for calling up nearby words and concepts. Those ideas, in turn, called up others as the model built up a sense of the situation. It composed its response with those ideas “in mind.”

A few months ago, I was reading an interview with an Anthropic researcher, Trenton Bricken, who has worked with colleagues to probe the insides of Claude, the company’s series of A.I. models. (Their research has not been peer-reviewed or published in a scientific journal.) His team has identified ensembles of artificial neurons, or “features,” that activate when Claude is about to say one thing or another. Features turn out to be like volume knobs for concepts; turn them up and the model will talk about little else. (In a sort of thought-control experiment, the feature representing the

Golden Gate Bridge was turned up; when one user asked Claude for a chocolate-cake recipe, its suggested ingredients included "1/4 cup dry fog" and "1 cup warm seawater.") In the interview, Bricken mentioned Google's Transformer architecture, a recipe for constructing neural networks that underlies leading A.I. models. (The "T" in ChatGPT stands for "Transformer.") He argued that the mathematics at the heart of the Transformer architecture closely approximated a model proposed decades earlier—by Pentti Kanerva, in "Sparse Distributed Memory."

Should we be surprised by the correspondence between A.I. and our own brains? L.L.M.s are, after all, artificial neural networks that psychologists and neuroscientists helped develop. What's more surprising is that when models practiced something rote—predicting words—they began to behave in such a brain-like way. These days, the fields of neuroscience and artificial intelligence are becoming entangled; brain experts are using A.I. as a kind of model organism. Evelina Fedorenko, a neuroscientist at M.I.T., has used L.L.M.s to study how brains process language. "I never thought I would be able to think about these kinds of things in my lifetime," she told me. "I never thought we'd have models that are good enough."

It has become commonplace to say that A.I. is a black box, but the opposite is arguably true: a scientist can probe the

activity of individual artificial neurons and even alter them. “Having a working system that instantiates a theory of human intelligence—it’s the dream of cognitive neuroscience,” Kenneth Norman, a Princeton neuroscientist, told me. Norman has created computer models of the hippocampus, the brain region where episodic memories are stored, but in the past they were so simple that he could only feed them crude approximations of what might enter a human mind. “Now you can give memory models the exact stimuli you give to a person,” he said.

The Wright brothers studied birds during their early efforts to build an airplane. They noted that birds take off into the wind, even though a reasonable person might have assumed they’d want the wind at their backs, and that they warp the tips of their wings for balance. These findings influenced their rudimentary glider designs. Then they built a six-foot-long wind tunnel, which allowed them to test a set of artificial wings under precisely controlled conditions. Their next round of glider flights was far more successful. Strangely, it was only well after they’d made a working flying machine that it became possible to understand exactly how the birds do it.

A.I. enables scientists to place thinking itself in a wind tunnel. For a paper provocatively titled “On the Biology of a Large Language Model,” Anthropic researchers observed Claude

responding to queries and described “circuits”—cascades of features that, together, perform complex computations. (Calling up the right memories is one step toward thinking; combining and manipulating them in circuits is arguably another.) One longstanding criticism of L.L.M.s has been that, because they must generate one token of their response at a time, they can’t plan or reason. But, when you ask Claude to finish a rhyming couplet in a poem, a circuit begins considering the last word of the new line, to insure that it will rhyme. It then works backward to compose the line as a whole. Anthropic researchers counted this as evidence that their models do engage in planning. Squint a little and you might feel, for the first time, that the inner workings of a mind are in view.

You really do have to squint, though. “The worry I have is that people flipped the bit from ‘I’m really skeptical of this’ to totally dropping their shields,” Norman, the Princeton neuroscientist, told me. “Many things still have to get figured out.” I’m one of the people that Norman is talking about. (Perhaps I am too easily moved by the seeming convergence of “Sparse Distributed Memory” and an Anthropic model.) In the past year or two, I started to believe what Geoffrey Hinton, who recently won a Nobel Prize for his A.I. research, told the journalist Karen Hao in 2020: “Deep learning is going to be able to do everything.” But we have also seen that larger models aren’t always better models. Curves

plotting model performance against size have begun flattening out. It's becoming difficult to find high-quality data that the models haven't already digested, and computing power is increasingly expensive. When GPT-5 came out, in August, it was a merely incremental improvement—and so profound a disappointment that it threatened to pop the A.I. investment bubble. The moment demands a middle kind of skepticism: one that takes today's A.I. models seriously without believing that there are no hard problems left.

Perhaps the most consequential of these problems is how to design a model that learns as efficiently as humans do. It is estimated that GPT-4 was exposed to trillions of words in training; children need only a few million to become fluent. Cognitive scientists tell us that a newborn's brain has certain "inductive biases" that accelerate learning. (Of course, the brain is the result of millions of years of evolution—itself a sort of training data.) For instance, human babies have the expectation that the world is made of objects, and that other beings have beliefs and intentions. When Mama says "banana," an infant connects that word to the entire yellow object she's looking at—not just its tip or its peel. Infants perform little experiments: Can I eat this? How far can I throw that? They are motivated by emotions such as desire, curiosity, and frustration. Children are always trying to do something just beyond their ability. Their learning is efficient because it's embodied, adaptive, deliberate, and continuous.

Maybe truly understanding the world requires participating in it.

An A.I.'s experience, in comparison, is so impoverished that it can't really be called "experience." Large language models are trained on data that is already extraordinarily refined. "I think the reason they work is that they're piggybacking on language," Tsao, the U.C. Berkeley neuroscientist, told me. Language is like experience pre-chewed; other kinds of data are less dense with meaning. "Why is it that we haven't had a comparable revolution in terms of reasoning about video data?" Gershman, the Harvard cognitive scientist, asked. "The kinds of vision models that we have still struggle with common-sense reasoning about physics." A recent model from DeepMind can generate videos in which paints are mixed correctly and mazes are solved—but they also depict a glass bouncing, instead of shattering, and ropes defying physics by being smooshed into a knot. Ida Momennejad, a cognitive neuroscientist who now works for Microsoft Research, has done experiments in which an L.L.M. is given a virtual walk-through of a building and then asked questions about routes and shortcuts—spatial inferences that come easily to humans. With all but the most basic setups, the A.I.s tend to fail or hallucinate nonexistent paths. "Do they really do planning?" she said. "Not really."

In my conversations with neuroscientists, I sensed a concern

that the A.I. industry is racing ahead somewhat thoughtlessly. If the goal is to make artificial minds as capable as human minds are, then "we're not training the systems in the right way," Brenden M. Lake, a cognitive scientist at Princeton, told me. When an A.I. is done training, the neural network's "brain" is frozen. If you tell the model some facts about yourself, it doesn't rewire its neurons. Instead, it uses a crude substitute: it writes down a bit of text —"The user has a toddler and is studying French"—and considers that before other instructions you give. The human brain updates itself continuously, and there's a beautiful theory about one of its ways of doing so: when you sleep, selected snapshots from your episodic memory are replayed for your neocortex in order to train it. Your high-dimensional thought space gets dimpled by the replayed memories; you wake up with a slightly new way of seeing.

The A.I. community has become so addicted to—and so financially invested in—breakneck progress that it sometimes pretends that advancement is inevitable and there's no science left to do. Science has the inconvenient property of sometimes stalling out. Silicon Valley may call A.I. companies "labs," and some employees there "researchers," but fundamentally it has an engineering culture that does whatever works. "It's just so remarkable how little the machine-learning community bothers looking at, let alone respects, the history and cognitive science that

precedes it," Cohen said.

Today's A.I. models owe their success to decades-old discoveries about the brain, but they are still deeply unlike brains. Which differences are incidental and which are fundamental? Every group of neuroscientists has its pet theory. These theories can be put to the test in a way that wasn't possible before. Still, no one expects easy answers. The problems that continue to plague A.I. models "are solved by carefully identifying ways in which the models don't behave as intelligently as we want them to and then addressing them," Norman said. "That is still a human-scientist-in-the-loop process."

In the nineties, billions of dollars poured into the Human Genome Project on the assumption that sequencing DNA might solve medicine's most vexing problems: cancer, hereditary conditions, even aging. It was a time of bluster and confidence—the era of Dolly the cloned sheep and "Jurassic Park"—when biotech was ascendant and the commentariat reckoned with whether humans should be playing God. Biologists soon found that the reality was more complicated. We didn't cure cancer or discover the causes of Alzheimer's or autism. We learned that DNA tells just one part of the story of life. In fact, one could argue that biology got swept up in a kind of gene fever, fixating on DNA because we had the means to study and understand it.

Still, nobody would claim that Francis Crick was wrong when, on the day in 1953 that he helped confirm the structure of DNA, he walked into a Cambridge pub talking about having discovered the secret of life. He and his colleagues did more to demystify life than almost anyone, ever. The decades following their discovery were among the most productive and exciting in the history of science. DNA became a household term; every high schooler learns about the double helix.

With A.I., we once again find ourselves in a moment of bluster and confidence. Sam Altman talks about raising half a trillion dollars to build Stargate, a new cluster of A.I. data centers, in the U.S. People discuss the race for superintelligence with a gravitas and an urgency that can seem ungrounded, even silly. But I suspect the reason that the Amodeis and Altmans of the world are making messianic pronouncements is that they believe that the basic picture of intelligence has been worked out; the rest is just details.

Even some neuroscientists believe that a crucial threshold has been crossed. “I really think it could be the right model for cognition,” Uri Hasson, a colleague of Cohen’s, Norman’s, and Lake’s at Princeton, said of neural networks. This upsets him as much as it excites him. “I have the opposite worry of most people,” he said. “My worry is not that these models are similar to us. It’s that we are similar to these models.” If

simple training techniques can enable a program to behave like a human, maybe humans aren't as special as we thought. Could it also mean that A.I. will surpass us not only in knowledge but also in judgment, ingenuity, cunning—and, as a result, power? To my surprise, Hasson told me that he is "worried these days that we might succeed in understanding how the brain works. Pursuing this question may have been a colossal mistake for humanity." He likened A.I. researchers to nuclear scientists in the nineteen-thirties: "This is the most interesting time in the life of these people. And, at the same time, they know that what they are working on has grave implications for humanity. But they cannot stop because of the curiosity to learn."

One of my favorite books by Hofstadter is a nerdy volume called "Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought." When I was in college, it electrified me. The premise was that a question such as "What is thinking?" was not merely philosophical but, rather, had a real answer. In 1995, when the book was published, Hofstadter and his research group could only gesture at what the answer might be. Thinking back on the book, I wondered whether Hofstadter would feel excited that A.I. researchers may have attained what he had yearned for: a mechanical account of the rudiments of thinking. When we spoke, however, he sounded profoundly disappointed—and frightened. Current A.I. research

“confirms a lot of my ideas, but it also takes away from the beauty of what humanity is,” he told me. “When I was younger, much younger, I wanted to know what underlay creativity, the mechanisms of creativity. That was a holy grail for me. But now I want it to remain a mystery.” Perhaps the secrets of thinking are simpler than anyone expected—the kind of thing that a high schooler, or even a machine, could understand. ♦