ARTIFICIAL INTELLIGENCE

Geoffrey Hinton tells us why he's now scared of the tech he helped build

"I have suddenly switched my views on whether these things are going to be more intelligent than us."

By Will Douglas Heaven

May 2, 2023

LINDA NYLIND / EYEVINE VIA REDUX

I met Geoffrey Hinton at his house on a pretty street in north London just four days before the bombshell announcement that he is quitting Google. Hinton is a pioneer of deep learning who helped develop some of the most important techniques at the heart of modern artificial intelligence, but after a decade at Google, he is stepping down to focus on new concerns he now has about AI.

Stunned by the capabilities of new large language models like GPT-4, Hinton wants to raise public awareness of the serious risks that he now believes may accompany the technology he ushered in.

he had plenty to say.

The 75-year-old computer scientist, who was a joint recipient with Yann LeCun and Yoshua Bengio of the 2018 Turing Award for his work on deep learning, says he is ready to shift gears. "I'm getting too old to do technical work that requires remembering lots of details," he told me. "I'm still okay, but I'm not nearly as good as I was, and that's annoying."

But that's not the only reason he's leaving Google. Hinton wants to spend his time on what he describes as "more philosophical work." And that will focus on the small but—to him—very real danger that AI will turn out to be a disaster.

Related Story

Deep learning pioneer Geoffrey Hinton has quit Google

Hinton will be speaking at EmTech Digital on Wednesday.

Leaving Google will let him speak his mind, without the self-censorship a Google executive must engage in. "I want to talk about AI safety issues without having to worry about how it interacts with Google's business," he says. "As long as I'm paid by Google, I can't do that."

That doesn't mean Hinton is unhappy with Google by any means. "It may surprise you," he says. "There's a lot of good things about Google that I want to say, and they're much more credible if I'm not at Google anymore."

Hinton says that the new generation of large language models—especially GPT-4, which OpenAI released in March—has made him realize that machines are on track to be a lot smarter than he thought they'd be. And he's scared about how that might play out.

"These things are totally different from us," he says. "Sometimes I think it's as if aliens had landed and people haven't realized because they speak very good English."

Foundations

Hinton is best known for his work on a technique called backpropagation, which he proposed (with a pair of colleagues) in the 1980s. In a nutshell, this is the algorithm that allows machines to learn. It underpins almost all neural networks today, from computer vision systems to large language models.

also trained a neural network to predict the next letters in a sentence, a precursor to today's large language models.

One of these graduate students was Ilya Sutskever, who went on to cofound OpenAI and lead the development of ChatGPT. "We got the first inklings that this stuff could be amazing," says Hinton. "But it's taken a long time to sink in that it needs to be done at a huge scale to be good." Back in the 1980s, neural networks were a joke. The dominant idea at the time, known as symbolic AI, was that intelligence involved processing symbols, such as words or numbers.

But Hinton wasn't convinced. He worked on neural networks, software abstractions of brains in which neurons and the connections between them are represented by code. By changing how those neurons are connected—changing the numbers used to represent them—the neural network can be rewired on the fly. In other words, it can be made to learn.

"My father was a biologist, so I was thinking in biological terms," says Hinton. "And symbolic reasoning is clearly not at the core of biological intelligence.

"Crows can solve puzzles, and they don't have language. They're not doing it by storing strings of symbols and manipulating them. They're doing it by changing the strengths of connections between neurons in their brain. And so it has to be possible to learn complicated things by changing the strengths of connections in an artificial neural network."

A new intelligence

For 40 years, Hinton has seen artificial neural networks as a poor attempt to mimic biological ones. Now he thinks that's changed: in trying to mimic what biological brains do, he thinks,

Hinton's fears will strike many as the stuff of science fiction. But here's his case.

As their name suggests, large language models are made from massive neural networks with vast numbers of connections. But they are tiny compared with the brain. "Our brains have 100 trillion connections," says Hinton. "Large language models have up to half a trillion, a trillion at most. Yet GPT-4 knows hundreds of times more than any one person does. So maybe it's actually got a much better learning algorithm than us."

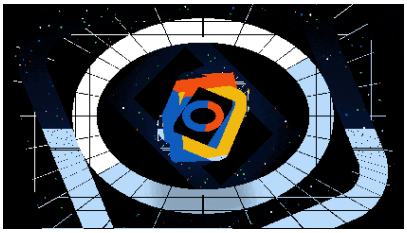
Compared with brains, neural networks are widely believed to be bad at learning: it takes vast amounts of data and energy to train them. Brains, on the other hand, pick up new ideas and skills quickly, using a fraction as much energy as neural networks do.

"People seemed to have some kind of magic," says Hinton. "Well, the bottom falls out of that argument as soon as you take one of these large language models and train it to do something new. It can learn new tasks extremely quickly."

Hinton is talking about "few-shot learning," in which pretrained neural networks, such as large language models, can be trained to do something new given just a few examples. For example, he notes that some of these language models can string a series of logical statements together into an argument even though they were never trained to do so directly.

Compare a pretrained large language model with a human in the speed of learning a task like that and the human's edge vanishes, he says.

Related Story



Geoffrey Hinton has a hunch about what's next for AI

A decade ago, the artificial-intelligence pioneer transformed the field with a major breakthrough. Now he's working on a new imaginary system named GLOM.

What about the fact that large language models make so much stuff up? Known as "hallucinations" by AI researchers (though Hinton prefers the term "confabulations," because it's the correct term in psychology), these errors are often seen as a fatal flaw in the technology. The tendency to generate them makes chatbots untrustworthy and, many argue, shows that these models have no true understanding of what they say.

Hinton has an answer for that too: bullshitting is a feature, not a bug. "People always confabulate," he says. Half-truths and misremembered details are hallmarks of human conversation: "Confabulation is a signature of human memory. These models are doing something just like people."

The difference is that humans usually confabulate more or less correctly, says Hinton. To Hinton, making stuff up isn't the problem. Computers just need a bit more practice.

most people have a hopelessly wrong view of how people work."

Of course, brains still do many things better than computers: drive a car, learn to walk, imagine the future. And brains do it on a cup of coffee and a slice of toast. "When biological intelligence was evolving, it didn't have access to a nuclear power station," he says.

But Hinton's point is that if we are willing to pay the higher costs of computing, there are crucial ways in which neural networks might beat biology at learning. (And it's worth pausing to consider what those costs entail in terms of energy and carbon.)

Learning is just the first string of Hinton's argument. The second is communicating. "If you or I learn something and want to transfer that knowledge to someone else, we can't just send them a copy," he says. "But I can have 10,000 neural networks, each having their own experiences, and any of them can share what they learn instantly. That's a huge difference. It's as if there were 10,000 of us, and as soon as one person learns something, all of us know it."

What does all this add up to? Hinton now thinks there are two types of intelligence in the world: animal brains and neural networks. "It's a completely different form of intelligence," he says. "A new and better form of intelligence."

That's a huge claim. But AI is a polarized field: it would be easy to find people who would laugh in his face—and others who would nod in agreement.

People are also divided on whether the consequences of this new form of intelligence, if it exists, would be beneficial or apocalyptic. "Whether you think superintelligence is going to be good or bad depends very much on whether you're an optimist or a pessimist," he says. "If you ask people to estimate the risks of bad things happening, like what's the chance of someone in your family getting really sick or being hit by a car, an optimist might say 5% and a pessimist might say it's guaranteed to happen. But the mildly depressed person will say the odds are maybe around 40%, and they're usually right."

Which is Hinton? "I'm mildly depressed," he says. "Which is why I'm scared."

How it could all go wrong

Hinton fears that these tools are capable of figuring out ways to manipulate or kill humans who aren't prepared for the new technology.

"I have suddenly switched my views on whether these things are going to be more intelligent than us. I think they're very close to it now and they will be much more intelligent than us in the future," he says. "How do we survive that?"

He is especially worried that people could harness the tools he himself helped breathe life into to tilt the scales of some of the most consequential human experiences, especially elections and wars.

"Look, here's one way it could all go wrong," he says. "We know that a lot of the people who want to use these tools are bad actors like Putin or DeSantis. They want to use them for winning wars or manipulating electorates."

Hinton believes that the next step for smart machines is the ability to create their own subgoals, interim steps required to carry out a task. What happens, he asks, when that ability is applied to something inherently immoral?

"Don't think for a moment that Putin wouldn't make hyper-intelligent robots with the goal of killing Ukrainians," he says. "He wouldn't hesitate. And if you want them to be good at it, you don't want to micromanage them—you want them to figure out how to do it."

There are already a handful of experimental projects, such as BabyAGI and AutoGPT, that hook chatbots up with other programs such as web browsers or word processors so that they can string together simple tasks. Tiny steps, for sure—but they signal the direction that some people want to take this tech. And even if a bad actor doesn't seize the machines, there are other concerns about subgoals, Hinton says.

"Well, here's a subgoal that almost always helps in biology: get more energy. So the first thing that could happen is these robots are going to say, 'Let's get more power. Let's reroute all the electricity to my chips.' Another great subgoal would be to make more copies of yourself. Does that sound good?"

Maybe not. But Yann LeCun, Meta's chief AI scientist, agrees with the premise but does not share Hinton's fears. "There is no question that machines will become smarter than humans—

machines will usher in a new renaissance for humanity, a new era of enlightenment," says LeCun. "I completely disagree with the idea that machines will dominate humans simply because they are smarter, let alone destroy humans."

"Even within the human species, the smartest among us are not the ones who are the most dominating," says LeCun. "And the most dominating are definitely not the smartest. We have numerous examples of that in politics and business."

Yoshua Bengio, who is a professor at the University of Montreal and scientific director of the Montreal Institute for Learning Algorithms, feels more agnostic. "I hear people who denigrate these fears, but I don't see any solid argument that would convince me that there are no risks of the magnitude that Geoff thinks about," he says. But fear is only useful if it kicks us into action, he says: "Excessive fear can be paralyzing, so we should try to keep the debates at a rational level."

Just look up

One of Hinton's priorities is to try to work with leaders in the technology industry to see if they can come together and agree on what the risks are and what to do about them. He thinks the international ban on chemical weapons might be one model of how to go about curbing the development and use of dangerous AI. "It wasn't foolproof, but on the whole people don't use chemical weapons," he says.

Related Story

AI pioneer Geoff Hinton: "Deep learning is going to be able to do everything"

Thirty years ago, Hinton's belief in neural networks was contrarian. Now it's hard to find anyone who disagrees, he says.

Bengio agrees with Hinton that these issues need to be addressed at a societal level as soon as possible. But he says the development of AI is accelerating faster than societies can keep up. The capabilities of this tech leap forward every few months; legislation, regulation, and international treaties take years.

This makes Bengio wonder whether the way our societies are currently organized—at both national and global levels—is up to the challenge. "I believe that we should be open to the possibility of fairly different models for the social organization of our planet," he says.

Does Hinton really think he can get enough people in power to share his concerns? He doesn't know. A few weeks ago, he watched the movie *Don't Look Up*, in which an asteroid zips toward Earth, nobody can agree what to do about it, and everyone dies—an allegory for how the world is failing to address climate change.

act when faced with serious threats. It is also true that AI risks causing real harm—upending the job market, entrenching inequality, worsening sexism and racism, and more. We need to focus on those problems. But I still can't make the jump from large language models to robot overlords. Perhaps I'm an optimist.

When Hinton saw me out, the spring day had turned gray and wet. "Enjoy yourself, because you may not have long left," he said. He chuckled and shut the door.

Be sure to tune in to Will Douglas Heaven's live interview with Hinton at EmTech Digital on Wednesday, May 3, at 1:30 Eastern time. Tickets are available from the event website.

by Will Douglas Heaven

DEEP DIVE

ARTIFICIAL INTELLIGENCE

ChatGPT is about to revolutionize the economy. We need to decide what that looks like.

New large language models will transform many jobs. Whether they will lead to widespread prosperity or not is up to us.

By David Rotman

ChatGPT is going to change education, not destroy it

The narrative around cheating students doesn't tell the whole story. Meet the teachers who think generative AI could actually make learning better.

By Will Douglas Heaven

GPT-4 is bigger and better than ChatGPT—but OpenAI won't say why

We got a first look at the much-anticipated big new language model from OpenAI. But this time how it works is even more deeply under wraps.

By Will Douglas Heaven

Deep learning pioneer Geoffrey Hinton has quit Google

Hinton will be speaking at EmTech Digital on Wednesday.

By Will Douglas Heaven

Illustration by Rose Wong

Contact us