IDEAS MADE TO MATTER | ARTIFICIAL INTELLIGENCE

Why neural net pioneer Geoffrey Hinton is sounding the alarm on AI

by Sara Brown | May 23, 2023

Why It Matters

Geoffrey Hinton, a respected researcher who recently stepped down from Google, said it's time to confront the existential dangers of artificial intelligence.

A deep learning pioneer is raising concerns about rapid advancements in artificial intelligence and how they will affect humans.

Geoffrey Hinton, 75, a professor emeritus at the University of Toronto and until recently a vice president and engineering fellow at Google, announced in early May that he was leaving the company — in part because of his age, he said, but also because he's changed his mind about the relationship between humans and digital intelligence.

In a widely discussed interview with The New York Times, Hinton said generative intelligence could spread misinformation and, eventually, threaten humanity.

Speaking two days after that article was published, Hinton reiterated his concerns. "I'm sounding the alarm, saying we have to worry about this," he said at the EmTech Digital conference, hosted by MIT Technology Review.

Hinton said he is worried about the increasingly powerful machines' ability to outperform humans in ways that are not in the best interest of humanity, and the likely inability to limit AI development.

The growing power of AI

In 2018, Hinton shared a Turing Award for work related to neural networks. He has been called "a godfather of AI," in part for his fundamental research about using back-propagation to help machines learn.

> *I think it's quite conceivable that humanity is just a passing phase in the evolution of intelligence.*
>
> **Geoffrey Hinton** | Former Vice President and Engineering Fellow, Google
>
> *Share*

Hinton said he long thought that computer models weren't as powerful as the human brain. Now, he sees artificial intelligence as a relatively imminent "existential threat."

Computer models are outperforming humans, including doing things humans can't do. Large language models like GPT-4 use neural networks with connections like those in the human brain and are starting to do commonsense reasoning, Hinton said.

These AI models have far fewer neural connections than humans do, but they manage to know a thousand times as much as a human, Hinton said.

In addition, models are able to continue learning and easily share knowledge. Many copies of the same AI model can run on different hardware but do exactly the same thing.

"Whenever one [model] learns anything, all the others know it," Hinton said. "People can't do that. If I learn a whole lot of stuff about quantum mechanics and I want you to know all that stuff about quantum mechanics, it's a long, painful process of getting you to understand it."

AI is also powerful because it can process vast quantities of data — much more than a single person can. And AI models can detect trends in data that aren't otherwise visible to a person — just like a doctor who had seen 100 million patients would notice more trends and have more insights than a doctor who had seen only a thousand.

AI concerns: Manipulating humans, or even replacing them

Hinton's concern with this burgeoning power centers around the alignment problem — how to ensure that AI is doing what humans want it to do. "What we want is some way of making sure that even if they're smarter than us, they're going to do things that are beneficial for us," Hinton said. "But we need to try and do that in a world where there [are] bad actors who want to build robot soldiers that kill people. And it seems very hard to me."

Humans have inherent motivations, such as finding food and shelter and staying alive, but AI doesn't. "My big worry is, sooner or later someone will wire into them the ability to create their own subgoals," Hinton said. (Some versions of the technology, like ChatGPT, already have the ability to do that, he noted.)

"I think it'll very quickly realize that getting more control is a very good subgoal because it helps you achieve other goals," Hinton said. "And if these things get carried away with getting more control, we're in trouble."

Artificial intelligence can also learn bad things — like how to manipulate people "by reading all the novels that ever were and everything Machiavelli ever wrote," for example. "And if [AI models] are much smarter than us, they'll be very good at manipulating us. You won't realize what's going on," Hinton said. "So even if they can't directly pull levers, they can certainly get us to pull levers. It turns out if you can manipulate people, you can invade a building in Washington without ever going there yourself."

At worst, "it's quite conceivable that humanity is just a passing phase in the evolution of intelligence," Hinton said. Biological intelligence evolved to create digital intelligence, which can absorb everything humans have created and start getting direct experience of the world.

"It may keep us around for a while to keep the power stations running, but after that, maybe not," he added. "We've figured out how to build beings that are immortal. These digital intelligences, when a piece of hardware dies, they don't die. If … you can find another piece of hardware that can run the same instructions, you can bring it to life again. So we've got immortality, but it's not for us."

Barriers to stopping AI advancement

Hinton said he does not see any clear or straightforward solutions. “I wish I had a nice, simple solution I could push, but I don’t,” he said. “But I think it’s very important that people get together and think hard about it and see whether there is a solution.”

More than 27,000 people, including several tech executives and researchers, have signed an open letter calling for a pause on training the most powerful AI systems for at least six months because of “profound risks to society and humanity,” and several leaders from the Association for the Advancement of Artificial Intelligence signed a letter calling for collaboration to address the promise and risks of AI.

RELATED ARTICLES

It’s not too late to rechart the course of technology

The AI road not taken

MIT Sloan research on AI and machine learning

It might be rational to stop developing artificial intelligence, but that’s naive and unlikely, Hinton said, in part because of competition between companies and countries.

“If you’re going to live in a capitalist system, you can’t stop Google [from] competing with Microsoft,” he said, noting that he doesn’t think Google, his former employer, has done anything wrong in developing AI programs. “It’s just inevitable in the capitalist system or a system with competition between countries like the U.S. and China that this stuff will be developed,” he said.

It is also hard to stop developing AI because there are benefits in fields like medicine, he noted.

Researchers are looking at guardrails for these systems, but there is the chance that AI can learn to write and execute programs itself. “Smart things can outsmart us,” Hinton said.

One note of hope: Everyone faces the same risk. “If we allow it to take over, it will be bad for all of us,” Hinton said. “We’re all in the same boat with respect to the existential threat. So we all ought to be able to cooperate on trying to stop it.”

STUDY: INDUSTRY NOW DOMINATES AI RESEARCH →

FOR MORE INFO

Sara Brown
Senior News Editor and Writer

sbrown1@mit.edu

FIND US

MIT Sloan School of Management
100 Main Street
Cambridge, MA 02142

617-253-1000