What Really Made Geoffrey Hinton Into an AI Doomer

The AI pioneer is alarmed by how clever the technology he helped create has become. And it all started with a joke.

Will Knight May 8, 2023 11:18 AM

Photograph: CHLOE ELLINGSON/Redux

Geoffrey Hinton, perhaps the most important person in the recent history of artificial intelligence, recently sent me a video of Snoop Dogg.

In the clip of a discussion panel, the rapper expresses profane amazement at how artificial intelligence software, such as ChatGPT, can now hold a coherent and meaningful conversation.

“Then I heard the old dude that created AI saying, ‘This is not safe ’cause the AIs got their own mind and these motherfuckers gonna start doing their own shit,’” Snoop says. “And I’m like, ‘Is we in a fucking movie right now or what?’”

The “old dude” is, of course, Hinton. He didn’t create AI exactly, but has played a major role in developing the artificial neural network foundations of today’s most powerful AI programs, including ChatGPT, the chatbot that has sparked widespread debate about how rapidly machine intelligence is progressing.

“Snoop gets it,” Hinton tells me over Zoom from his home in London. The researcher recently left Google so that he could more freely call attention to

the risks posed by intelligent machines. Hinton says AI is advancing more quickly than he and other experts expected, meaning there is an urgent need to ensure that humanity can contain and manage it. He is most concerned about near-term risks such as more sophisticated, AI-generated disinformation campaigns, but he also believes the long-term problems could be so serious that we need to start worrying about them now.

When asked what triggered his newfound alarm about the technology he has spent his life working on, Hinton points to two recent flashes of insight.

One was a revelatory interaction with a powerful new AI system—in his case, Google's AI language model PaLM, which is similar to the model behind ChatGPT, and which the company made accessible via an API in March. A few months ago, Hinton says he asked the model to explain a joke that he had just made up—he doesn't recall the specific quip—and was astonished to get a response that clearly explained what made it funny. "I'd been telling people for years that it's gonna be a long time before AI can tell you why jokes are funny," he says. "It was a kind of litmus test."

Hinton's second sobering realization was that his previous belief that software needed to become much more complex—akin to the human brain—to become significantly more capable was probably wrong. PaLM is a large program, but its complexity pales in comparison to the brain's, and yet it could perform the kind of reasoning that humans take a lifetime to attain.

Hinton concluded that as AI algorithms become larger, they might outstrip their human creators within a few years. "I used to think it would be 30 to 50 years from now," he says. "Now I think it's more likely to be five to 20."

Hinton isn't the only person to have been shaken by the new capabilities that large language models such as PaLM or GPT-4 have begun demonstrating. Last month, a number of prominent AI researchers and others signed an

open letter calling for a pause on the development of anything more powerful than currently exists. But since leaving Google, Hinton feels his views on whether the development of AI should continue have been misconstrued.

“A lot of the headlines have been saying that I think it should be stopped now—and I've never said that,” he says. “First of all, I don't think that's possible, and I think we should continue to develop it because it could do wonderful things. But we should put equal effort into mitigating or preventing the possible bad consequences.”

Hinton says he didn’t leave Google to protest its handling of this new form of AI. In fact, he says, the company moved relatively cautiously despite having a lead in the area. Researchers at Google invented a type of neural network known as a transformer, which has been crucial to the development of models like PaLM and GPT-4.

In the 1980s, Hinton, a professor at the University of Toronto, along with a handful of other researchers, sought to give computers greater intelligence by training artificial neural networks with data instead of programming them in the conventional way. The networks could digest pixels as input, and, as they saw more examples, adjust the values connecting their crudely simulated neurons until the system could recognize the contents of an image. The approach showed fits of promise over the years, but it wasn’t until a decade ago that its real power and potential became apparent.

In 2018, Hinton was given the Turing Award, the most prestigious prize in computer science, for his work on neural networks. He received the prize together with two other pioneering figures, Yann LeCun, Meta’s chief AI scientist, and Yoshua Bengio, a professor at the University of Montreal.

That’s when a new generation of many-layered artificial neural networks—

fed copious amounts of training data and run on powerful computer chips—were suddenly far better than any existing program at labeling the contents of photographs.

The technique, known as deep learning, kicked off a renaissance in artificial intelligence, with Big Tech companies rushing to recruit AI experts, build increasingly powerful deep learning algorithms, and apply them to products such as face recognition, translation, and speech recognition.

Google hired Hinton in 2013 after acquiring his company, DNNResearch, founded to commercialize his university lab's deep learning ideas. Two years later, one of Hinton's grad students who had also joined Google, Ilya Sutskever, left the search company to cofound OpenAI as a nonprofit counterweight to the power being amassed by Big Tech companies in AI.

Since its inception, OpenAI has focused on scaling up the size of neural networks, the volume of data they guzzle, and the computer power they consume. In 2019, the company reorganized as a for-profit corporation with outside investors, and later took $10 billion from Microsoft. It has developed a series of strikingly fluent text-generation systems, most recently GPT-4, which powers the premium version of ChatGPT and has stunned researchers with its ability to perform tasks that seem to require reasoning and common sense.

Hinton believes we already have a technology that will be disruptive and destabilizing. He points to the risk, as others have done, that more advanced language algorithms will be able to wage more sophisticated misinformation campaigns and interfere in elections.

The most impressive new capabilities of GPT-4 and models like PaLM are what he finds most unsettling. The fact that AI models can perform complex logical reasoning and interact with humans, and are progressing more

quickly than expected, leads some to worry that we are getting closer to seeing algorithms capable of outsmarting humans seeking more control. "What really worries me is that you have to create subgoals in order to be efficient, and a very sensible subgoal for more or less anything you want to do is to get more power—get more control," Hinton says.

Some of those raising the alarm about AI have been extreme in their claims. Eliezer Yudkowsky, a researcher at the nonprofit Machine Intelligence Research Institute, has claimed in a recent TED talk, as well as in an article for *Time*, that AI is on course to kill everyone on earth and that nations should be willing to use deadly force to ensure the development of AI comes to a stop. "I listened to him thinking he was going to be crazy. I don't think he's crazy at all," Hinton says. "But, okay, it's not helpful to talk about bombing data centers."

Recent leaps in AI also conjure up utopian ideas. Hinton points to Ray Kurzweil, another AI pioneer now at Google. "Ray wants to be immortal," he says. "Well, the good news is we've figured out how to make immortal beings, the bad news is it's not for us. But can you imagine if all old white men hung around forever?"

But Hinton also confesses that he doesn't know how to control the AI that OpenAI, Google, and others are building. "I really don't know," he says. "All I'm saying is a lot of smart people should be putting a lot of effort into figuring out how we deal with the possibility of AI taking over as one of all the other possibilities."

Hinton certainly believes that AI scientists now have a vital role in drawing attention to the risks that may lie ahead, devising new safeguards, and working across international lines. "Maybe I should actually be talking to Chinese scientists," he says, and suggests he might send an email to Andrew

Yao, a professor at Tsinghua University in Beijing who, like him, won the Turing Award and is famous for his research in AI.

I ask Hinton whether he sees the effort to mitigate the emerging risks posed by AI as a kind of Manhattan Project, which would perhaps make him a modern J. Robert Oppenheimer. “They just had to make something go bang, but it is a lot harder to make sure something doesn’t,” he says.

Despite the importance of his warning, Hinton has not lost his acerbic sense of humor, as is clear when he explains why a more advanced form of AI would inevitably become unruly, even dangerous.

“How many examples do you know of a more intelligent thing being controlled by a less intelligent thing—well, since Biden got elected of course,” he says. “Oh, and you can quote me on that last bit.”