‘We’ve discovered the secret of immortality. The bad news is it’s not for us’: why the godfather of AI fears for humanity

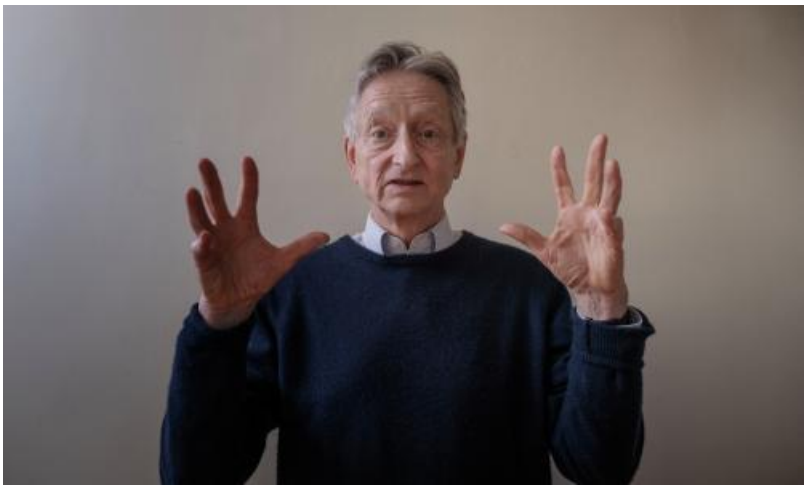
Publication Date: 2023-05-05

Author: Alex Hern

Section: Technology

Tags: Artificial intelligence (AI), Google, Computing, interviews

Article URL: https://www.theguardian.com/technology/2023/may/05/geoffrey-hinton-godfather-of-ai-fears-for-humanity



The first thing Geoffrey Hinton says when we start talking, and the last thing he repeats before I turn off my recorder, is that he left Google, his employer of the past decade, on good terms. “I have no objection to what Google has done or is doing, but obviously the media would love to spin me as ‘a disgruntled Google employee’. It’s not like that.” It’s an important clarification to make, because it’s easy to conclude the opposite. After all, when most people calmly describe their former employer as being one of a small group of companies charting a course that is alarmingly likely to wipe out humanity itself, they do so with a sense of opprobrium. But to listen to Hinton, we’re about to sleepwalk towards an existential threat to civilisation without anyone involved acting maliciously at all. Known as one of three “godfathers of AI”, in 2018 Hinton won the ACM Turing award – the Nobel prize of computer scientists for his work on “deep learning”. A cognitive psychologist and computer scientist by training, he wasn’t motivated by a desire to radically improve technology: instead, it was to understand more about ourselves. “For the last 50 years, I’ve been trying to make computer models that can learn stuff a bit like the way the brain learns it, in order to understand better how the brain is learning things,” he tells me when we meet in his sister’s house in north London, where he is staying (he usually resides in Canada). Looming slightly over me – he prefers to talk standing up, he says – the tone is uncannily reminiscent of a university tutorial, as the 75-year-old former professor explains his research history, and how it has inescapably led him to the conclusion that we may be doomed. In trying to model how the human brain works, Hinton found himself one of the leaders in the field of “neural networking”, an approach to building computer systems that can learn from data and experience. Until recently, neural nets were a curiosity, requiring vast computer power to perform simple tasks worse than other approaches. But in the last decade, as the availability of processing power and vast datasets has exploded, the approach Hinton pioneered has ended up at the centre of a technological revolution. “In trying to think about how the brain could implement the algorithm behind all these models, I decided that maybe it can’t – and maybe these big models are actually much better than the brain,” he says. A “biological intelligence” such as ours, he says, has advantages. It runs at low power, “just 30 watts, even when you’re thinking”, and “every brain is a bit different”. That means we learn by mimicking others. But that approach is “very inefficient” in terms of information transfer. Digital intelligences, by contrast, have an enormous advantage: it’s trivial to share information between multiple copies. “You pay an enormous cost in terms of energy, but when one of them learns something, all of them know it, and you can easily store more copies. So the good news is, we’ve discovered the secret of immortality. The bad news is, it’s not for us.”

Once he accepted that we were building intelligences with the potential to outthink humanity, the more alarming conclusions followed. “I thought it would happen eventually, but we had plenty of time: 30 to 50 years. I don’t think that any more. And I don’t know any examples of more intelligent things being controlled by less intelligent things – at least, not since Biden got elected. “You need to imagine something more intelligent than us by the same difference that we’re more intelligent than a frog. And it’s going to learn from the web, it’s going to have read every single book that’s ever been written on how to manipulate people, and also seen it in practice.” He now thinks the crunch time will come in the next five to 20 years, he says. “But I wouldn’t rule out a year or two. And I still wouldn’t rule out 100 years – it’s just that my confidence that this wasn’t coming for quite a while has been shaken by the realisation that biological intelligence and digital intelligence are very different, and digital intelligence is probably much better.” There’s still hope, of sorts, that AI’s potential could prove to be over-stated. “I’ve got huge uncertainty at present. It is possible that large language models,” the technology that underpins systems such as ChatGPT, “having consumed all the documents on the web, won’t be able to go much further unless they can get access to all our private data as well. I don’t want to rule things like that out – I think people who are confident in this situation are crazy.” Nonetheless, he says, the right way to think about the odds of disaster is closer to a simple coin toss than we might like. This development, he argues, is an unavoidable consequence of technology under capitalism. “It’s not that Google’s been bad. In fact, Google is the leader in this research, the core technical breakthroughs that underlie this wave came from Google, and it decided not to release them directly to the public. Google was worried about all the things we worry about, it has a good reputation and doesn’t want to mess it up. And I think that was a fair, responsible decision. But the problem is, in a capitalist system, if your competitor then does do that, there’s nothing you can do but do the same.” He decided to quit his job at Google, he has said, for three reasons. One was simply his age: at 75, he’s “not as good at the technical stuff as I used to be, and it’s very annoying not being as good as you used to be. So I decided it was time to retire from doing real work.” But rather than remain in a nicely remunerated ceremonial position, he felt it was important to cut ties entirely, because, “if you’re employed by a company, there’s inevitable self-censorship. If I’m employed by Google, I need to keep thinking, ‘How is this going to impact Google’s business?’ And the other reason is that there’s actually a lot of good things I’d like to say about Google, and they’re more credible if I’m not at Google.” Since going public about his fears, Hinton has come under fire for not following some of his colleagues in quitting earlier. In 2020, Timnit Gebru, the technical co-lead of Google’s ethical AI team, was fired by the company after a dispute over a research paper spiralled into a wide-ranging clash over the company’s diversity and inclusion policies. A letter signed by more than 1,200 Google staffers opposed the firing, saying it “heralds danger for people working for ethical and just AI across Google”. But there is a split within the AI faction over which risks are more pressing. “We are in a time of great uncertainty,” Hinton says, “and it might well be that it would be best not to talk about the existential risks at all so as not to distract from these other things [such as issues of AI ethics and justice]. But then, what if because we didn’t talk about it, it happens?” Simply focusing on the short-term use of AI, to solve the ethical and justice issues present in the technology today, won’t necessarily improve humanity’s chances of survival at large, he says. Not that he knows what will. “I’m not a policy guy. I’m just someone who’s suddenly become aware that there’s a danger of something really bad happening. I want all the best brains who know about AI – not just philosophers, politicians and policy wonks but people who actually understand the details of what’s happening – to think hard about these issues. And many of them are, but I think it’s something we need to focus on.” Since he first spoke out on Monday, he’s been turning down requests from the world’s media at a rate of one every two minutes (he agreed to meet with the Guardian, he said, because he has been a reader for the past 60 years, since he switched from the Daily Worker in the 60s). “I have three people who currently want to talk to me – Bernie Sanders, Chuck Schumer and Elon Musk. Oh, and the White House. I’m putting them all off until I have a bit more time. I thought when I retired I’d have plenty of time to myself.” Throughout our conversation, his lightly jovial tone of voice is somewhat at odds with the message of doom and destruction he’s delivering. I ask him if he has any reason for hope. “Quite often, people seem to come out of situations that appeared hopeless, and be OK. Like, nuclear weapons: the cold war with these powerful weapons seemed like a very bad situation. Another example would be the ‘Year 2000’ problem. It was nothing like this existential risk, but the fact that people saw it ahead of time and made a big fuss about it meant that people overreacted, which was a lot better than under-reacting. “The reason it was never a problem is because people actually sorted it out before it happened.”