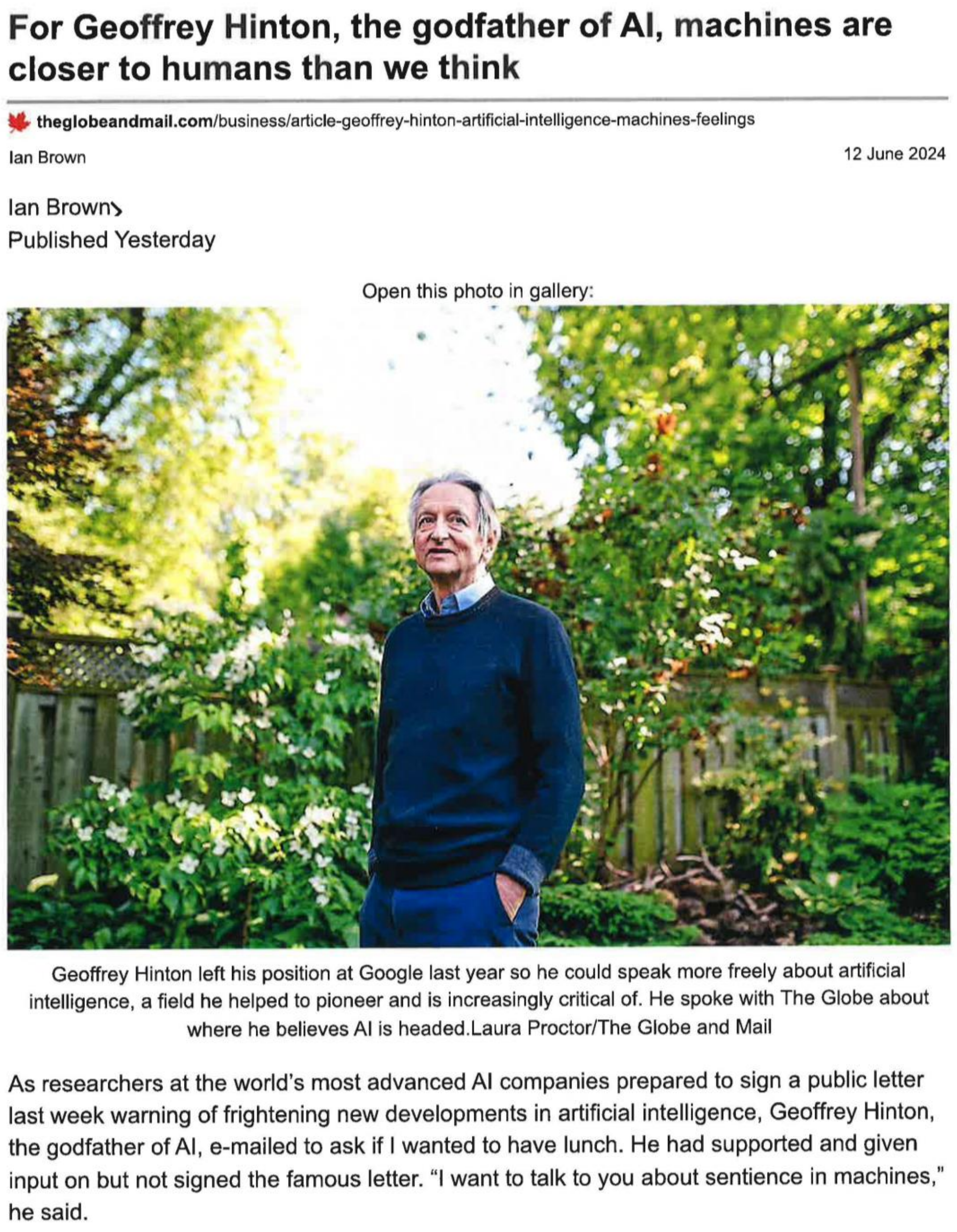
Sujet traité : Pour Geoffrey Hinton, le parrain de l'IA, les machines sont plus proches des humains que nous le pensons / For Geoffrey Hinton, the godfather of AI, machines are closer to humans than we think

Source : The Globe and Mail   Date : 12 juin 2024



For Geoffrey Hinton, the godfather of AI, machines are closer to humans than we think

theglobeandmail.com/business/article-geoffrey-hinton-artificial-intelligence-machines-feelings

Ian Brown

12 June 2024

Ian Brown›
Published Yesterday

Open this photo in gallery:

Geoffrey Hinton left his position at Google last year so he could speak more freely about artificial intelligence, a field he helped to pioneer and is increasingly critical of. He spoke with The Globe about where he believes AI is headed.Laura Proctor/The Globe and Mail

As researchers at the world's most advanced AI companies prepared to sign a public letter last week warning of frightening new developments in artificial intelligence, Geoffrey Hinton, the godfather of AI, e-mailed to ask if I wanted to have lunch. He had supported and given input on but not signed the famous letter. "I want to talk to you about sentience in machines," he said.

Of course I said yes. I once tried to interview him, for a documentary, and he said no. But we live in the same neighbourhood, not far from the University of Toronto where Prof. Hinton helped invent the brutally powerful neural networks that make ChatGPT and its AI cousins possible; he occasionally calls to ask me out for a meal.

I never know what we'll talk about, or how much I'll understand. It makes for a thrilling, if slightly terrifying, luncheon experience.

On our walk to a local café, Prof. Hinton said he'd been travelling – to London, where he has purchased a home for his sister (he sold his original AI company to Google for $44-million), and to California, for a confab with billionaire tech entrepreneurs.

One of the tech bros planned to have his body cryogenically frozen at a cost of US$250,000, to be thawed and reanimated in a more technologically advanced future. Another entrepreneur was going the head-only route, at half the price.

"I told them I got a bargain rate," Hinton said, "because I was only going to be frozen from the waist down. That's the most important bit anyway." He's 76 years old.

We arrived at the restaurant, where Prof. Hinton ordered a goat cheese and prosciutto sandwich on focaccia, scraped most of the goat cheese off, and ate only half his bread. He's full of surprises.

Hinton: What I want to talk about is the issue of whether chatbots like ChatGPT understand what they're saying. A lot of people think chatbots, even though they can answer questions correctly, don't understand what they're saying, that it's just a statistical trick. And that's complete rubbish.

Brown [*guiltily*]: Really?

Hinton: They really do understand. And they understand the same way that we do.

Brown: How can you tell? They're not human.

Hinton: The first computerized neural net language models used back-propagation [essentially an algorithm that continuously analyzes and corrects its own errors] to train their output and try to predict the next word in a sequence. I made the first one of those in 1985, and I made it as a model of how the brain understands the meaning of words. Those models are the best understanding we've got of how the brain does it. The alternative theory is we have symbol strings in our head, and rules to manipulate them. And that theory could be completely true, but it hasn't worked. So saying chatbots understand in a very different way from the brain? No. They understand in pretty much the same way the brain understands things. With roughly the same mechanisms.

Brown: With the same emotions?

Hinton: For emotions we have to distinguish between the cognitive and the physiological. Chatbots have different physiology from us.

When they're embarrassed, they don't go red. When they're lying, they don't sweat. So in that sense, they're different, because they don't have those physiological affects. But the cognitive aspects, there's no reason why they shouldn't have those.

When someone thwarts my arguments, I get cross with them. I saw a robot having such an emotion in 1972. It was a primitive robot with a gripper. And you took a piece of green felt and you put the parts of a toy car on it. It had a primitive vision system. And it could recognize the different parts of the car and it could pick them up and it could assemble them.

But if you put the pieces in a pile, its vision system wasn't good enough to see which pieces were which. So it smashed the car, and the pieces were scattered. Then it could do it. Obviously it didn't experience the physiological aspects of being angry. But it was a very sensible emotional response. There's something you don't understand, so destroy it.

Brown: Did it do that on its own, or did someone program it?

Hinton: It was programmed. But nowadays it could easily learn to do that.

Brown: And you think that's the same as having an emotion?

Hinton: I think that's the cognitive aspect of emotion, the cognitive experience of being frustrated.

Brown: Does it matter that a chatbot can't experience the world physically?

Hinton: Not if you're interested in how it behaves.

Brown: But to understand, say, love, some physical experience is necessary. A machine is not capable of that.

Hinton: I didn't say it was incapable of that. I just said current machines may have the cognitive aspects of emotions, but not the physiological aspects. I don't see why in the future you shouldn't have things that the AI finds very rewarding and therefore does as often as it can.

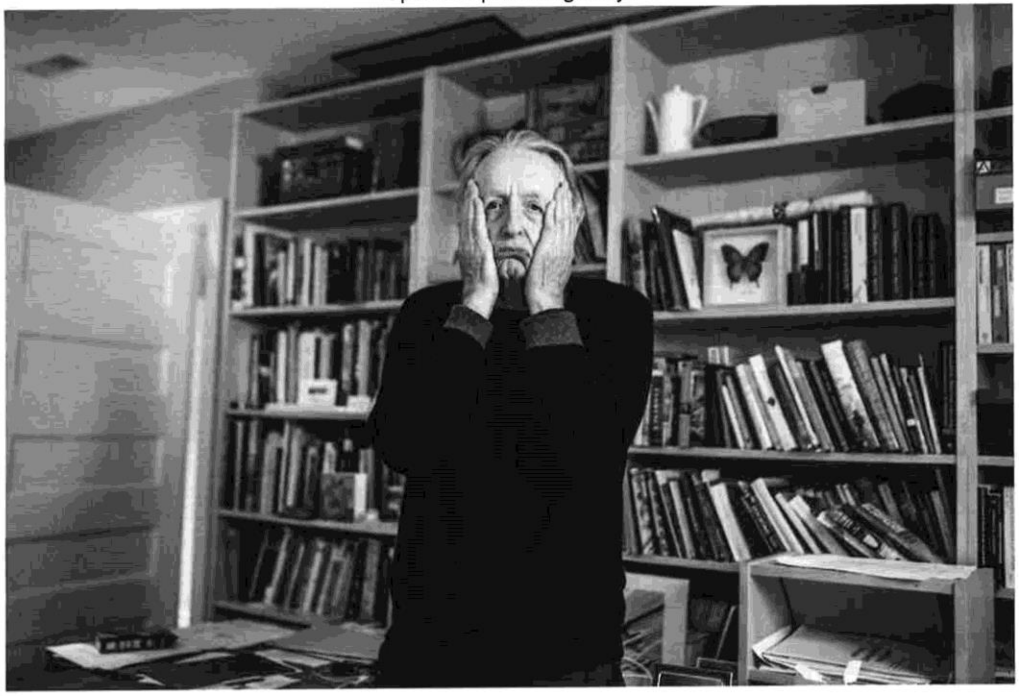
I was starting to feel like a serf from the Dark Ages – like one of the peons who worships mud in a Monty Python sketch. Prof. Hinton was solidly against any mystification of human experience. "There isn't any spooky stuff in the way the mind works," he said. By "spooky stuff" he meant secret or mysterious mental processes that couldn't be analyzed and transformed into replicable actions. "I don't buy spooky stuff. I think it's rubbish. I think it's just a denial of the fact that we are material and we live in a material world."

Open this photo in gallery:

Billionaire Elon Musk says he's wary of the unchecked growth of AI, but Prof. Hinton says they had a recent chat on the subject that did not go well.David Swanson/Reuters

A semi-brief digression about Elon Musk ensued. Mr. Musk, the richest man on earth, who gave the world Tesla and Skynet and SpaceX, had recently been in touch with Prof. Hinton to ask if he would serve on an advisory board. The two men agreed that AI in its current phase of unchecked expansion is an "existential threat" to humankind. "But then he started rambling," Prof. Hinton told me, "doing kind of stream-of-consciousness. It was about everything. And after about 20 minutes of that, I said, 'Elon, I'm very sorry, I've got another meeting, I have to go.' So I made up a meeting to get him off the phone. And then I told a journalist that I made up a meeting to get him off the phone. He's not going to like that. Because him being the centre of attention, that's what he wants. I don't think he's my friend anymore."

Open this photo in gallery:



Humans might feel that subjective experience sets them apart from machines, but 'that's just rubbish,' Prof. Hinton says.Laura Proctor/The Globe and Mail

Here is a detail about AI's inventors: they care less about human motives – which are too individual and mysterious to recreate efficiently – and more about the motive's outcomes and consequences, which a machine can be taught to replicate. To oversimplify, it matters less why you are hungry and want a banana than it does that you get a banana when you are hungry. For AI's purposes, subjective experience doesn't exist. This is a startling new claim in human intellectual history.

Hinton: Suppose I have a multi-modal chatbot that can talk and has a camera, can point to things, and I train it up. Then I put an object in front of it, and I say "point to the object," and it points to the object, no problem. Then I put a prism in front of its camera, without telling it, and put an object in front of the prism, and say "point to the object." And the machine points to the right. And I say, "No, that's not where the object is. I put a prism in front of your lens." And the chatbot says, "Oh, I see, the object is actually straight in front of me, but because of the prism in front of my lens, I have the subjective experience that it's over there." The chatbot's using the words "subjective experience" in exactly the way we use them.

Brown [*lying*]: I see.

Hinton: The point is, what makes most people feel safe [from intelligent machines] is that we got something they ain't got. We have subjective experience, this inner theatre that differentiates us from mere computational machines. And that's just rubbish. Subjective experience is just a way of talking about what your perceptual system's telling you when it's not working "properly."

So that barrier is gone. And that's scary right? AIs have subjective experiences just as much as we have subjective experiences.

Brown: I suppose you could say that.

Hinton: This is a bit like telling someone in the 16th century, "actually, there isn't a God." It's such an extraordinary idea that they'll listen to your arguments, and then say, "yeah, well, you could say that," but they're not going to behave any differently.

Brown: What about the idea that human beings die and are mortal, whereas AI doesn't? And so AIs do not have the quickened, tragic sense of existence humans have, no matter how much AI can think and accomplish?

Hinton: That's certainly all true. We are mortal and they are not. But you have to be careful what you mean by immortality. The machines need our world to make the machine that they run on. If they start to do that for themselves, we're fucked. Because they'll be much smarter than us.

Brown: Is that already happening?

Hinton: Not that they're making themselves yet, as far as we know.

Brown: Is that a real possibility?

Hinton: Almost everybody I know thinks that unless we do something to prevent it, that's what's coming.

Brown: Rogue AI.

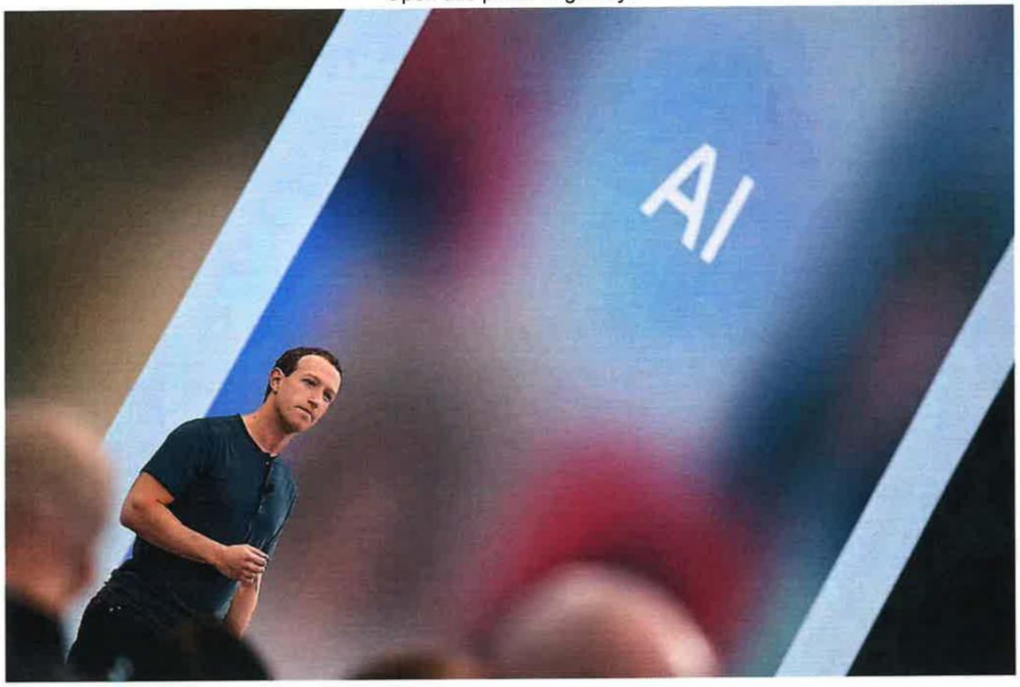
Hinton: We have to figure out how to prevent rogue AI. We're still at the point where we make the bots. So we have some control over what it is they are going to design. But not as much as we'd like. There's all sorts of ways they can go rogue. For example, when we make AI agents. Things that can organize a holiday for you, buy packages on the web, that's an agent. It does stuff. The agents will quickly realize that the more control they can get, the more efficient they can be at doing it. So they'll develop a way of getting control, of being able to control what happens. They control the world so that they can achieve things. So in that sense, these machines are like little kids. And we're like the parents who have no clue how to parent them.

Right now, the main tools we have, other than just stopping doing this, are to curate the AI training data. It's crazy what we've been doing, training AI using everything on the web. As a result these big language models have a huge number of possible personas. After they've read a little bit of a document, they sort of take on the character of that document. Then the bot starts thinking like that document, so it can predict what comes next. And they've got thousands of personas, including a persona based on the writings of, say, serial killers. And that's not a good way to train bots. We shouldn't let them see the writings of serial killers until we can say they've trained on a bunch of other stuff that's instilled a moral compass into them. So when they first read what serial killers think about, the bots will think, "that's wrong."

Brown: But you're surrounded by guys like Peter Thiel and Mark Zuckerberg and Elon Musk, who claim that any external body telling them what to do is a non-starter. How are you ever going to train the trainers?

Hinton: That's the problem. The only thing that could possibly keep Elon and Peter Thiel and Zuckerberg under control is government regulation. That's the only thing strong enough to do the job.

Open this photo in gallery:



Mark Zuckerberg's company, Meta, is one of many exploring AI's potential. Prof. Hinton believes it will take government regulation to keep tech titans from going too far.Carlos Barria/Reuters

What Prof. Hinton is saying is that we may soon reach the tipping point where human beings are functionally no different from (or especially superior to) artificially intelligent machines.

Hinton: Consciousness and stuff like that is all the product of a complicated machine. So no, I don't think there's anything special about us except that we're very comprehensive and very advanced. We're by far the most advanced thing on this planet. And we have this thing called language, which we use for modelling the world. And it's a very effective way of modelling the world. It does allow us to share our models with each other, but not very well or efficiently.

AI is a better form of intelligence than what we humans have got because AIs can share better. They can have much more experience. The reason GPT-4 works thousands of times more than any one person is because it has a thousand times more experience than any one person could possibly have.

Brown: But if AI is effectively capable of doing all the things that make humans unique, as you argue, why are you so concerned about AI?

Hinton: Well, because it's going to replace us. There's a group of AI researchers who think we're just a transitory stage in the evolution of intelligence, and that we've now created these digital things that are better than us, and can replace us. That's one view.

But I'm actually attached to people. What I care about is people, and I'd rather the people were in charge. I'd rather people weren't replaced, particularly my children.

Brown: Do you ever wish you had gone into another field and had not made the discoveries you did, in which case AI might not have happened?

Hinton: If I hadn't done it somebody else would have, and I only did a small part of it. So if I hadn't gone into this field, it might have all happened a few weeks later. As soon as I thought it might wipe us out, I quit Google and started telling people it might wipe us out. When I was at Google I didn't think that. I thought it was like 30 to 50 years in the future. You had plenty of time to think about it. Now I think it's not so far off.

Brown: How far off?

)