Magazine: 06 May 2023

Ben Lazarus

The godfather of AI: why I left Google

From magazine issue: 06 May 2023

Text / Comments

Ten minutes before I meet Geoffrey Hinton, the 'godfather of AI', the *New York Times* announces he's leaving Google. After decades working on artificial intelligence, Hinton now believes it could wipe out humanity. 'It is like aliens have landed on our planet and we haven't quite realised it yet because they speak very good English,' he says. He also tells me that he has been unable to sleep for months.

'It's conceivable that the genie is already out of the bottle'

Hinton, 75, revolutionised AI not once but twice: first with his work on neural networks, computer architecture that closely resembles the brain's structure, and then with so-called 'deep learning', which allows AI to refine and extract patterns and concepts on a vast quantity of data. He persevered with neural networks during the 1970s and 1980s when the industry had largely abandoned it. In 1986 he designed the very first chatbot, then called a 'language predictor model'. Decades later, in 2012, he developed a deep-learning AI, the intellectual foundation for ChatGPT.

Hinton comes from a family of mathematicians and scientists (his grandfather was George Boole, who formulated Boolean algebra, laying the foundations for the entire digital age). He has had his qualms about AI in the past, but only over its application rather than the fundamentals. He loathed the development of autonomous weapons, and left his US job because of the State Department funding of AI. He's also been worried about the ability of AI to help 'bad actors' such as Vladimir Putin pump out fake news. 'I'd always kind of ignored the existential threat before about it wiping out humanity,' he says. 'Now I think there's a significant chance of that.'

What changed? 'Google had a big language model called Palm which could explain why a joke was funny. That shocked me. I arrived at the conclusion – this might just be a much better form of intelligence. If it is, it'll replace us.'

How likely is it that billions of people could be wiped out by AI? 'So much of all of this is unknown territory… I don't want to give a number,' he says, pacing from the

kitchen table to the oven at his home in north London. I push him for specifics. 'I am going to summarise in a very sloppy way what the French philosopher Blaise Pascal said: if you have two alternatives [i.e. AI killing humanity or not] and you have very little understanding of what is going on, then you should say 50 per cent.'

The next AI landmark will be Artificial General Intelligence (AGI): AI that is at least as competent as a human thinker. 'It's the first time in mankind's history that we're on the brink of developing something smarter than us,' he says. 'It still uses hugely more power, but it's comparable in intelligence now. Not as good, but getting there. And many people think it's going to get smarter than us soon. Not in a hundred years and not in 50 years – maybe 20 years, maybe five years. It may be that we're going to reach a new stage of evolution when biological intelligence gets replaced by digital intelligence. We may be history.'

Hinton, who won the Turing Award in 2018 (often called the 'Nobel Prize of Computing'), asks me to think of 'cases where a more intelligent thing is controlled by a less intelligent thing'. After a few embarrassing 'ums', he rescues me. 'There's Covid. That virus isn't that intelligent. It's not making decisions about what should happen – it's just screwing with us. It's not in charge.

'The only example I know of a more intelligent thing being controlled by a less intelligent thing is a mother and baby. But the mother is wired so that she can't stand the baby crying. Also, she cares a lot about the baby's welfare. And there's not as much difference in intelligence there: these things are going to be much more intelligent than us.'

With the tech giants (Microsoft, Google et al) thrashing it out to be top dog in the field, Hinton says that 'it's conceivable that the genie is already out of the bottle'.

Yet he is defensive of Google (he sold his company to the tech giant in 2012 for $44 million and opened Google's AI lab in Toronto) and blames Microsoft for forcing a market war. 'Google had the lead in AI for years. Google behaved very responsibly by not releasing chatbots. It had them quite a long time ago… But you can only do that when you're the leader. As soon as OpenAI gave ChatGPT to Microsoft – who funded the computing required – then Google had to put these things out there. There was no choice.'

The chief scientist at OpenAI, Ilya Sutskever, is a former student of Hinton's. 'I talk to him regularly,' Hinton says. 'He's quite concerned. But he doesn't think that the problems will be solved by OpenAI stopping research.'

Hinton is also pessimistic about whether any brakes can be applied, because of the competition between the US and China. 'I don't think there's any chance whatsoever of getting a pause. If they paused in the West, China wouldn't pause. There's a race going on.'

Does Hinton regret his lifetime's work? 'I don't really have regrets. There's the standard excuse if I hadn't done it, somebody else would have,' he says. 'But also, until very recently it looked like the benefits were going to be great and were certain; the risks were a long way off and not certain. The decision to work on it may have turned out to be unfortunate, but it was a wise decision at the time.'

His reasons for leaving Google are 'complicated': 'I'm 75. And I'm finding it harder and harder to do the technical work, because you have to remember all sorts of fine details and that's tricky. Another reason is I want to actually be able to tell the public how responsible Google has been so far. I want to say good things about Google, and it's much more credible if I'm not there. The third reason is, even if Google doesn't tell you what you should and shouldn't say, there's inevitable self-censorship if you work for an organisation… I'm aware of self-censorship. And so I don't want to be constrained by it. I just want to be able to say what I believe.'

At the heart of his Oppenheimer-style U-turn is the fear that the human brain isn't as impressive as digital intelligence. 'It was always assumed before that the brain was the best thing there was, and the things that we were producing were kind of wimpy attempts to mimic the brain. For 49 of the 50 years I was working on it, that's what I believed – brains were better.'

The human brain, he explains, runs on very low power: 'about 30 watts and we've got about 100 trillion connections'. He says trying to learn knowledge from someone 'is a slow and painful business'.

Digital intelligence requires much more energy but is shared across entire networks. 'If we fabricate it accurately, then you can have thousands and thousands of agents. When one agent learns something, all the others know it instantly… They can process so much more data than we can and see all sorts of things we'll never see.

‘A way of thinking about this is: if you go to your doctor with a rare condition, you’re lucky if they have seen even one case before. Now imagine going to a doctor who’d seen 100 million patients, including dozens who have this rare condition.’ So there are ‘wonderful’ short-term gains for AI being used in medicine and elsewhere, he says. ‘Anywhere where humans use intelligence, it’s going to help.’ Then he adds, with a smile: ‘In particular, it’s going to help where it’s a kind of not very acute intelligence like law.’

How exactly does Hinton think this digital intelligence could harm humanity? ‘Imagine it [AI] has the power to perform actions in the world as opposed to just answering questions. So a little bit of that power would be the power to connect to the internet and look things up on the internet, which chatbots didn’t originally have. Now imagine a household robot where you could tell it what to do and it can do things. That household robot will be a lot smarter than you. Are you confident it would keep doing what you told it to? We don’t know the answer. It’s kind of like the sorcerer’s apprentice.’

One of the big concerns for Hinton is that as AI progresses it will develop sub-goals to work more efficiently in achieving its main goal, but these sub-goals won’t necessarily align with human objectives and that will make us vulnerable to AI manipulation.

‘If you look at a baby in a highchair, its mother gives him a spoon to feed itself. And what’s the first thing he does? He drops it on the floor and the mother picks it up. He looks at his mother and drops it again.

‘The baby is trying to get control of his mother. There’s a very good reason for that – the more control you have, the easier it is to achieve your other goals. That’s why, in general, having power is good, because it allows you to achieve other things… Inevitably people will give these systems the ability to create sub-goals, because that’s how to make them efficient. One of the sub-goals they will immediately derive is to get more power, because that makes everything else easier.’

Even if AI manipulated us to have sub-goals to be more efficient, would they necessarily want more power? Here’s where evolution comes in, according to Hinton.

‘A sort of basic principle of evolution is if you take two varieties of a species, the one that produces more viable offspring wins. It’s going to end up replacing the other one. You see it operating very fast with viruses, like Omicron. The virus with a higher infection rate wins. But that’s true for all species. That’s how evolution works. Things that can produce more viable offspring win.

‘One of the sub-goals the systems will derive is to get more power, because that makes everything else easier’

‘Now imagine there are several different AGI. Which one’s going to win? The one that produces more copies of itself. So what worries me is if AGI ever got the idea that it should produce lots of copies of itself, then the one that was best at doing that would wipe out the others. It’s not clear that being nice to humans is going to help it produce more copies of itself. I don’t see why we shouldn’t get evolution among AGI.’

Even worse for us is that these machines can’t die. ‘If one of those digital computers dies, it doesn’t matter. You haven’t lost the knowledge. Also, if you just record the knowledge in a memory medium, then as soon as you have another digital computer, you can download it – it’s alive again.’

But surely we could just unplug a dodgy AI and deprive it of electricity – would that not stop them? Hinton smiles at me and paraphrases Hal from *2001: A Space Odyssey*. ‘I’m sorry, Dave. I can’t answer that question.’

WRITTEN BY

Ben Lazarus

Ben Lazarus is special projects editor of The Spectator

TOPICS Society

Most popular

1 *Andrew Tettenborn*
The university union may be beyond redemption

2 *Charles Parton*
The next Chinese tech threat is already here

3 *Annabel Denham*
Starmer's economic promises would spell disaster for the UK

4 *Gavin Mortimer*
France's failure to tackle migration is a warning to the Tories

5 *Steerpike*
Why won't Humza Yousaf condemn Celtic fans?

Read next

TRENDING

Gavin Mortimer

France's failure to tackle migration is a warning to the Tories

ALSO IN **SOCIETY**

Gareth Roberts

Succession's only real flaw

ALSO BY **BEN LAZARUS**

Ben Lazarus

Bare and spectral: Bob Dylan's Fragments – Time Out Of Mind Sessions reviewed

From Spectator Life

LATEST

David Loyn

Why Iran and the Taliban are clashing over water

Comments

ON OFF