TechXplore

Scientists warn of AI dangers but don't agree on solutions

May 3 2023, by Matt O'brien

Computer scientist Geoffrey Hinton poses at Google's Mountain View, Calif, headquarters on Wednesday, March 25, 2015. Computer scientists who helped build the foundations of today's artificial intelligence technology are warning of its dangers, but that doesn't mean they agree on the risks or how to prevent disastrous outcomes. Credit: AP Photo/Noah Berger, File

Computer scientists who helped build the foundations of today's artificial intelligence technology are warning of its dangers, but that doesn't mean they agree on what those dangers are or how to prevent them.

Humanity's survival is threatened when "smart things can outsmart us," so-called Godfather of AI Geoffrey Hinton said at a conference Wednesday at the Massachusetts Institute of Technology.

"It may keep us around for a while to keep the power stations running," Hinton said. "But after that, maybe not."

After retiring from Google so he could speak more freely, the 75-year-old Hinton said he's recently changed his views about the reasoning capabilities of the computer systems he's spent a lifetime researching.

"These things will have learned from us, by reading all the novels that ever were and everything Machiavelli ever wrote, how to manipulate people," Hinton said, addressing the crowd attending MIT Technology Review's EmTech Digital conference from his home via video. "Even if they can't directly pull levers, they can certainly get us to pull levers."

"I wish I had a nice simple solution I could push, but I don't," he added. "I'm not sure there is a solution."

Fellow AI pioneer Yoshua Bengio, co-winner with Hinton of the top computer science prize, told The Associated Press on Wednesday that he's "pretty much aligned" with Hinton's concerns brought on by chatbots such as ChatGPT and related technology, but worries that to simply say "We're doomed" is not going to help.

"The main difference, I would say, is he's kind of a pessimistic person, and I'm more on the optimistic side," said Bengio, a professor at the University of Montreal. "I do think that the dangers—the short-term ones, the long-term ones—are very serious and need to be taken seriously by not just a few researchers but governments and the population."

Aidan Gomez, co-founder and CEO Cohere, is photographed at an office in Palo Alto, Calif., Wednesday, April 12, 2023. Gomez, whose research behind the so-called “transformer” technique put the “T” at the end of ChatGPT, says some fearmongering is “detached from the reality” of AI’s true capabilities, but he’s among a growing number of experts pushing for safeguards.Credit: AP Photo/Jeff Chiu

There are plenty of signs that governments are listening. The White House has called in the CEOs of Google, Microsoft and ChatGPT-maker

OpenAI to meet Thursday with Vice President Kamala Harris in what's being described by officials as a frank discussion on how to mitigate both the near-term and long-term risks of their technology. European lawmakers are also accelerating negotiations to pass sweeping new AI rules.

But all the talk of the most dire future dangers has some worried that hype around superhuman machines—which don't exist—is distracting from attempts to set practical safeguards on current AI products that are largely unregulated and have been shown to cause real-world harms.

Margaret Mitchell, a former leader on Google's AI ethics team, said she's upset that Hinton didn't speak out during his decade in a position of power at Google, especially after the 2020 ouster of prominent Black scientist Timnit Gebru, who had studied the harms of large language models before they were widely commercialized into products such as ChatGPT and Google's Bard.

"It's a privilege that he gets to jump from the realities of the propagation of discrimination now, the propagation of hate language, the toxicity and nonconsensual pornography of women, all of these issues that are actively harming people who are marginalized in tech," said Mitchell, who was also forced out of Google in the aftermath of Gebru's departure. "He's skipping over all of those things to worry about something farther off."

Bengio, Hinton and a third researcher, Yann LeCun, who works at Facebook parent Meta, were all awarded the Turing Prize in 2019 for their breakthroughs in the field of artificial neural networks, instrumental to the development of today's AI applications such as ChatGPT.

Bengio, the only one of the three who didn't take a job with a tech giant, has voiced concerns for years about near-term AI risks, including job market destabilization, automated weaponry and the dangers of biased data sets.

But those concerns have grown recently, leading Bengio to join other computer scientists and tech business leaders like Elon Musk and Apple co-founder Steve Wozniak in calling for a six-month pause on developing AI systems more powerful than OpenAI's latest model, GPT-4.

Bengio said Wednesday he believes the latest AI language models already pass the "Turing test" named after British codebreaker and AI pioneer Alan Turing's method introduced in 1950 to measure when AI becomes indistinguishable from a human—at least on the surface.

Aidan Gomez, co-founder and CEO Cohere, is photographed at an office in Palo Alto, Calif., Wednesday, April 12, 2023. Gomez, whose research behind the so-called “transformer” technique put the “T” at the end of ChatGPT, says some fearmongering is “detached from the reality” of AI’s true capabilities, but he’s among a growing number of experts pushing for safeguards. Credit: AP Photo/Jeff Chiu

"That's a milestone that can have drastic consequences if we're not careful," Bengio said. "My main concern is how they can be exploited for nefarious purposes to destabilize democracies, for cyberattacks, disinformation. You can have a conversation with these systems and think that you're interacting with a human. They're difficult to spot."

Where researchers are less likely to agree is on how current AI language systems—which have many limitations, including a tendency to fabricate information—might actually get smarter than humans not just in memorizing huge troves of information, but in showing critical reasoning and other human skills.

Aidan Gomez was one of the co-authors of the pioneering 2017 paper that introduced a so-called transformer technique—the "T" at the end of ChatGPT—for improving the performance of machine-learning systems, especially in how they learn from passages of text. Then just a 20-year-old intern at Google, Gomez remembers laying on a couch at the company's California headquarters when his team sent out the paper around 3 a.m. when it was due.

"Aidan, this is going to be so huge," he remembers a colleague telling him, of the work that's since helped lead to new systems that can generate humanlike prose and imagery.

Six years later and now CEO of his own AI company called Cohere, which Hinton has invested in, Gomez is enthused about the potential applications of these systems but bothered by fearmongering he says is "detached from the reality" of their true capabilities and "relies on extraordinary leaps of imagination and reasoning."

"The notion that these models are somehow gonna get access to our nuclear weapons and launch some sort of extinction-level event is not a productive discourse to have," Gomez said. "It's harmful to those real pragmatic policy efforts that are trying to do something good."

Asked about his investments in Cohere on Wednesday in light of his broader concerns about AI, Hinton said he had no plans to pull his investments because there are still many helpful applications of language models in medicine and elsewhere. He also said he hadn't made any bad decisions in pursuing the research he started in the 1970s.

"Until very recently, I thought this existential crisis was a long way off," Hinton said. "So I don't really have any regrets about what I did."

Citation: Scientists warn of AI dangers but don't agree on solutions (2023, May 3) retrieved 18 April 2024 from https://techxplore.com/news/2023-05-scientists-ai-dangers-dont-solutions.html