AI firms ‘should include members of public on boards to protect society’

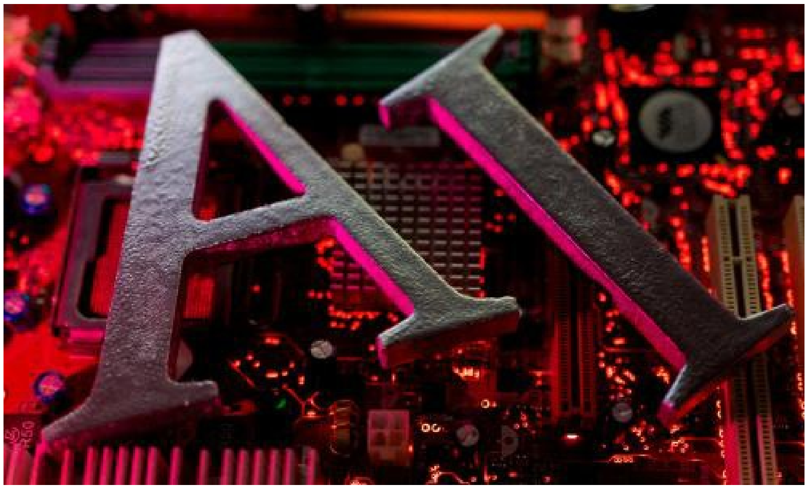
Publication Date: 2023-12-06

Author: Dan Milmo

Section: Technology

Tags: Artificial intelligence (AI), OpenAI, news

Article URL: https://www.theguardian.com/technology/2023/dec/06/ai-firms-should-include-members-of-public-on-boards-to-protect-society



Companies developing powerful artificial intelligence systems must have independent board members representing the “interests of society”, according to an expert regarded as one of the modern godfathers of the technology. Yoshua Bengio, a co-winner of the 2018 Turing Award – referred to as the “Nobel prize of computing” – said AI firms must have oversight from members of the public, as advances in the technology accelerate rapidly. Speaking in the wake of the boardroom upheaval at the ChatGPT developer OpenAI, including the exit and return of its chief executive, Sam Altman, Bengio said a “democratic process” was needed to monitor developments in the field. “How do we make sure that these advances are happening in a way that doesn’t endanger the public? How do we make sure that they’re not abused for increasing one’s power?” the AI pioneer told the Guardian. “To me, the answer is obvious in principle. We need democratic governance. We need an inclusive set of people on the board of these organisations, who can have visibility on what is going on and can act in a way different from the regulators. We need regulators but we also need people inside who are independent of the company and represent the interests of society.” Bengio, a full professor at the University of Montreal and the founder and scientific director of Mila, the Quebec Artificial Intelligence Institute, said he was concerned that despite recent turmoil at OpenAI there would not be a slowdown in AI development across the tech industry. In March, Bengio signed an open letter along with thousands of senior tech figures, including Elon Musk, calling for a six-month pause in development of the most powerful AI systems. “My concern is … that there’s not going to be a slowdown,” he said. “There’s going to be just racing ahead without the proper guardrails and with maybe much more of a focus on competition against the other players and winning that game, than protecting the public and safety of the public.” Soon after the reinstatement of Altman as the OpenAI CEO, days after he had been sacked, reports emerged that the company had been working on an AI model before his ousting whose capabilities had concerned some company researchers. Concerns over AI safety range from mass-produced disinformation to biased outcomes and accelerated development of artificial general intelligence, the term for a system that can carry out human tasks at human or above human levels of intelligence and potentially evade human control. “Anything that accelerates the timeline towards AGI is something that we should be concerned about, at least until we have the right guardrails, which I think are still missing,” Bengio said. The new-look OpenAI board comprises independent members and is chaired by Bret Taylor, the former chair of Twitter, and has retained one member of the board that fired Altman, the tech entrepreneur Adam D’Angelo. However, another independent board member, the AI safety researcher Helen Toner, who had co-authored a paper raising concerns about the impact of ChatGPT’s release on the speed of AI development elsewhere, has left.

Bengio added that he had concerns about a voluntary agreement between governments and AI companies at last month’s UK-hosted global AI safety summit to cooperate on testing powerful AI models before and after their deployment. He said the process “favours companies” because it requires governments to find problems with the models rather than putting the burden on companies to prove that their technology is safe. Bengio added that the tests under the framework announced at the Bletchley Park gathering would be “just spot checks”. “I think a much better recipe is that companies should have the burden to demonstrate to the regulator that their system can be trusted. Just like we ask pharmaceutical companies to do clinical studies [on their products],” he said. “It’s not the government doing the clinical studies, it’s the pharma [industry]. And then they have to come up with scientific evidence, like a statistical evaluation that says, ‘with very high probability drug is not going to be toxic’. So the government looks at that report and the process and says, yes, you can go ahead and commercialise it.” Bengio, who attended the Bletchley summit, was announced as the chair of the inaugural “state of AI science” report, which is expected to be published before a further AI summit in Korea in May. He said the report would be “very focused” on safety and expressed hope that it would be published every six months. “Hopefully there’ll be one every six months because the technology moves pretty fast,” he said. Bengio said the potential timeline for the emergence of a system that could evade human control was between five and 20 years. “My personal timeline for losing control is something more like five years at least, maybe 20. But I have a lot of uncertainty. It could happen faster. And if you’re the government, you should protect the public against this 1% odd possibility that something bad could happen with those systems,” he said. He also welcomed Joe Biden’s executive order on AI, published shortly before the Bletchley summit, describing it as a “very good thing” that moved the industry “towards better regulation”, although other countries needed to follow suit. The White House order’s provisions include requiring tech companies to share test results from the most powerful AI systems with the US government. Bengio won the 2018 Turing Award, along with Geoffrey Hinton and Yann LeCun, for conceptual and engineering breakthroughs that have helped make deep neural networks a critical component of computing.