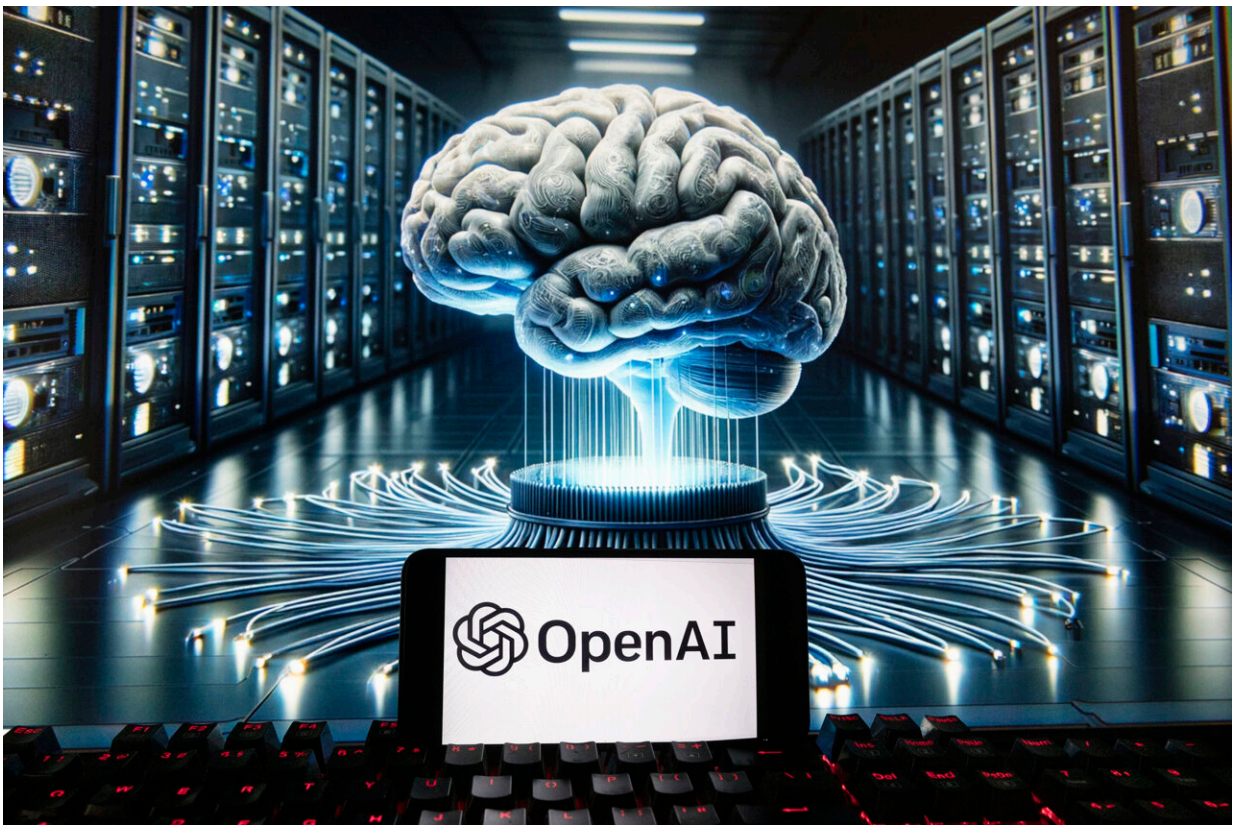
Former OpenAI employees lead push to protect whistleblowers flagging artificial intelligence risks

June 4 2024, by Matt O'brien



The OpenAI logo is seen displayed on a cell phone with an image on a computer monitor generated by ChatGPT's Dall-E text-to-image model, Dec. 8, 2023, in Boston. A group of OpenAI's current and former workers is calling on the ChatGPT-maker and other artificial intelligence companies to protect whistleblowing employees who flag safety risks about AI technology.Credit: AP Photo/Michael Dwyer, File

A group of OpenAI's current and former workers is calling on the ChatGPT-maker and other artificial intelligence companies to protect employees who flag safety risks about AI technology.

An open letter published Tuesday asks tech companies to establish stronger whistleblower protections so researchers have the "right to warn" about AI dangers without fear of retaliation.

The development of more powerful AI systems is "moving fast and there are a lot of strong incentives to barrel ahead without adequate caution," said former OpenAI engineer Daniel Ziegler, one of the organizers behind the open letter.

Ziegler said in an interview Tuesday he didn't fear speaking out internally during his time at OpenAI between 2018 to 2021, in which he helped develop some of the techniques that would later make ChatGPT so successful. But he now worries that the race to rapidly commercialize the technology is putting pressure on OpenAI and its competitors to disregard the risks.

Another co-organizer, Daniel Kokotajlo, said he quit OpenAI earlier this year "because I lost hope that they would act responsibly," particularly as it attempts to build better-than-human AI systems known as artificial general intelligence.

"They and others have bought into the 'move fast and break things' approach and that is the opposite of what is needed for technology this powerful and this poorly understood," Kokotajlo said in a written statement.

OpenAI said in response to the letter that it already has measures for

employees to express concerns, including an anonymous integrity hotline.

"We're proud of our track record providing the most capable and safest AI systems and believe in our scientific approach to addressing risk," said the company's statement. "We agree that rigorous debate is crucial given the significance of this technology and we'll continue to engage with governments, civil society and other communities around the world."

The letter has 13 signatories, most of whom are former employees of OpenAI and two who work or worked for Google's DeepMind. Four are listed as anonymous current employees of OpenAI. The letter asks that companies stop making workers enter into "non-disparagement" agreements that can punish them by taking away a key financial perk—their equity investments—if they criticize the company after they leave.

Social media outrage over language in OpenAI's paperwork for departing workers recently led the company to release all its former employees from those agreements.

The open letter has the support of pioneering AI scientists Yoshua Bengio and Geoffrey Hinton, who together won computer science's highest award, and Stuart Russell. All three have warned about the risks that future AI systems could pose to humanity's existence.

The letter comes as OpenAI has said it is beginning to develop the next generation of the AI technology behind ChatGPT. It formed a new safety committee just after losing a set of leaders, including co-founder Ilya Sutskever, who were part of a team focused on safely developing the most powerful AI systems.

The broader AI research community has long battled over the gravity of AI's short-term and long-term risks and how to square them with the technology's commercialization. Those conflicts contributed to the ouster, and swift return, of OpenAI CEO Sam Altman last year, and continue to fuel distrust in his leadership.

More recently, a new product showcase drew the ire of Hollywood star Scarlett Johansson, who said she was shocked to hear ChatGPT's voice sounding "eerily similar" to her own despite having previously rejected Altman's request that she lend her voice to the system.

Several who signed the letter, including Ziegler, have ties to effective altruism, a philanthropic social movement that embraces causes such as mitigating the potential worst impacts of AI. Ziegler said it's not just "catastrophic" future risks of out-of-control AI systems that the letter's authors are concerned about, but also fairness, product misuse, job displacement and the potential for highly realistic AI to manipulate people without the right safeguards.

"I'm less interested in scolding OpenAI," he said. "I am more interested in this being an opportunity for all the frontier AI companies to do something that really increases oversight and transparency and maybe increases public trust."

Citation: Former OpenAI employees lead push to protect whistleblowers flagging artificial intelligence risks (2024, June 4) retrieved 26 June 2024 from https://techxplore.com/news/2024-06-openai-employees-whistleblowers-flagging-artificial.html