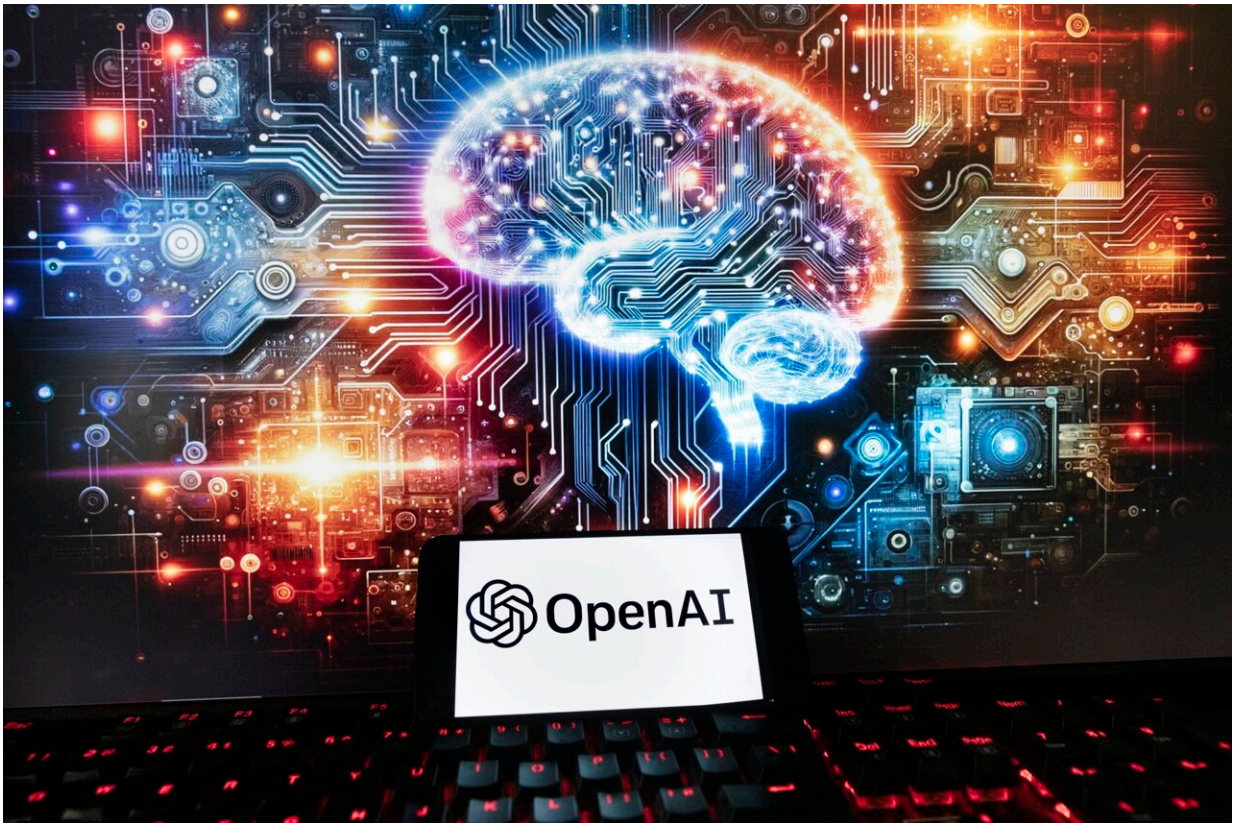
OpenAI forms safety committee as it starts training latest artificial intelligence model

May 28 2024, by Associated Press



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, Friday, Dec. 8, 2023, in Boston. OpenAI says it's setting up a new safety and security committee and has begun training a new artificial intelligence model to supplant the GPT-4 system that underpins its ChatGPT chatbot. The San Francisco startup said in a blog post Tuesday May 28, 2024 that the committee will advise the full board on “critical safety and security decisions" for its projects and operations. Credit: AP Photo/Michael Dwyer, File

OpenAI says it's setting up a safety and security committee and has begun training a new AI model to supplant the GPT-4 system that underpins its ChatGPT chatbot.

The San Francisco startup said in a blog post Tuesday that the committee will advise the full board on "critical safety and security decisions" for its projects and operations.

The safety committee arrives as debate swirls around AI safety at the company, which was thrust into the spotlight after a researcher, Jan Leike, resigned and leveled criticism at OpenAI for letting safety "take a backseat to shiny products." OpenAI co-founder and chief scientist Ilya Sutskever also resigned, and the company disbanded the "superalignment" team focused on AI risks that they jointly led.

Leike said Tuesday he's joining rival AI company Anthropic, founded by ex-OpenAI leaders, to "continue the superalignment mission" there.

OpenAI said it has "recently begun training its next frontier model" and its AI models lead the industry on capability and safety, though it made no mention of the controversy. "We welcome a robust debate at this important moment," the company said.

AI models are prediction systems that are trained on vast datasets to generate on-demand text, images, video and human-like conversation. Frontier models are the most powerful, cutting edge AI systems.

The safety committee is filled with company insiders, including OpenAI CEO Sam Altman and Chairman Bret Taylor, and four OpenAI technical and policy experts. It also includes board members Adam D'Angelo, who's the CEO of Quora, and Nicole Seligman, a former Sony general

counsel.

The committee's first job will be to evaluate and further develop OpenAI's processes and safeguards and make its recommendations to the board in 90 days. The company said it will then publicly release the recommendations it's adopting "in a manner that is consistent with safety and security."