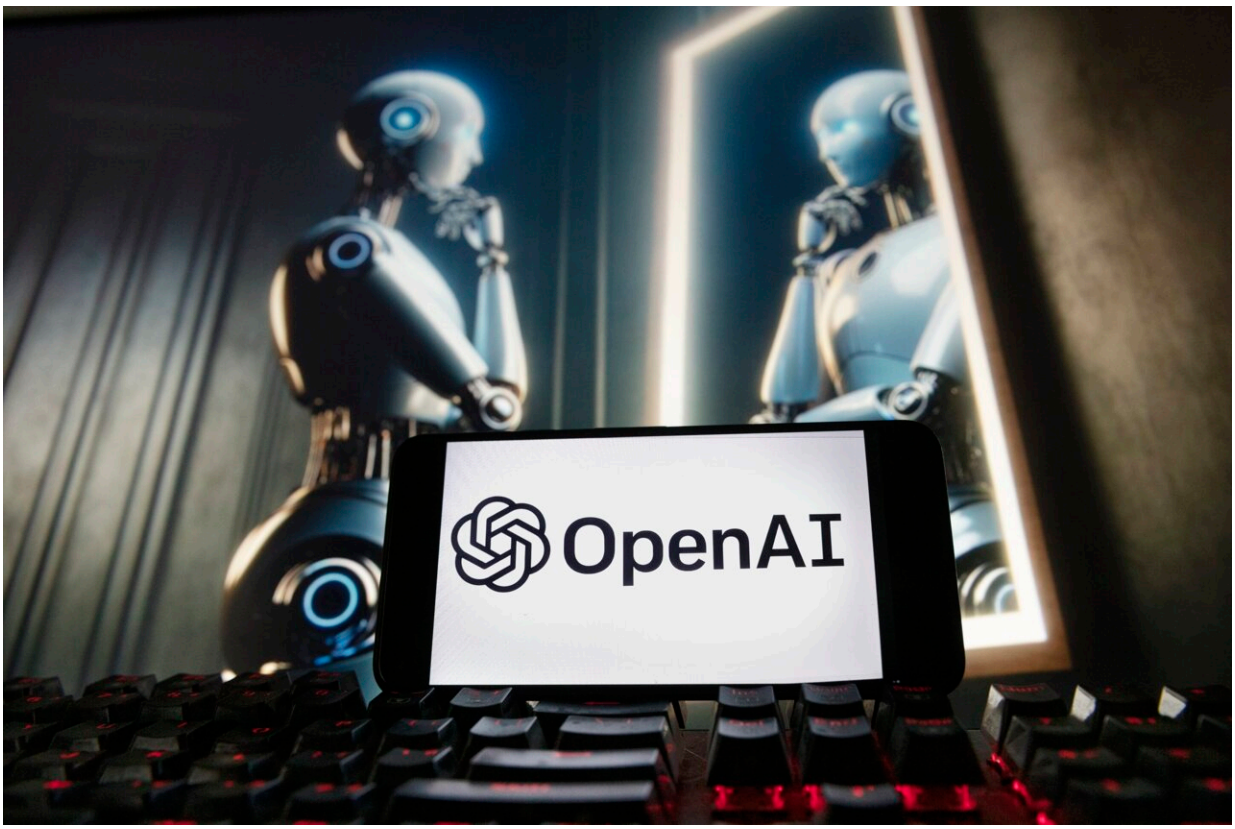
A former OpenAI leader says safety has 'taken a backseat to shiny products' at the AI company

May 18 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. A former OpenAI leader who resigned from the company earlier this week said on Friday that product safety has "taken a backseat to shiny products" at the influential artificial intelligence company. Credit: AP Photo/Michael Dwyer, file

A former OpenAI leader who resigned from the company earlier this week said Friday that safety has "taken a backseat to shiny products" at the influential artificial intelligence company.

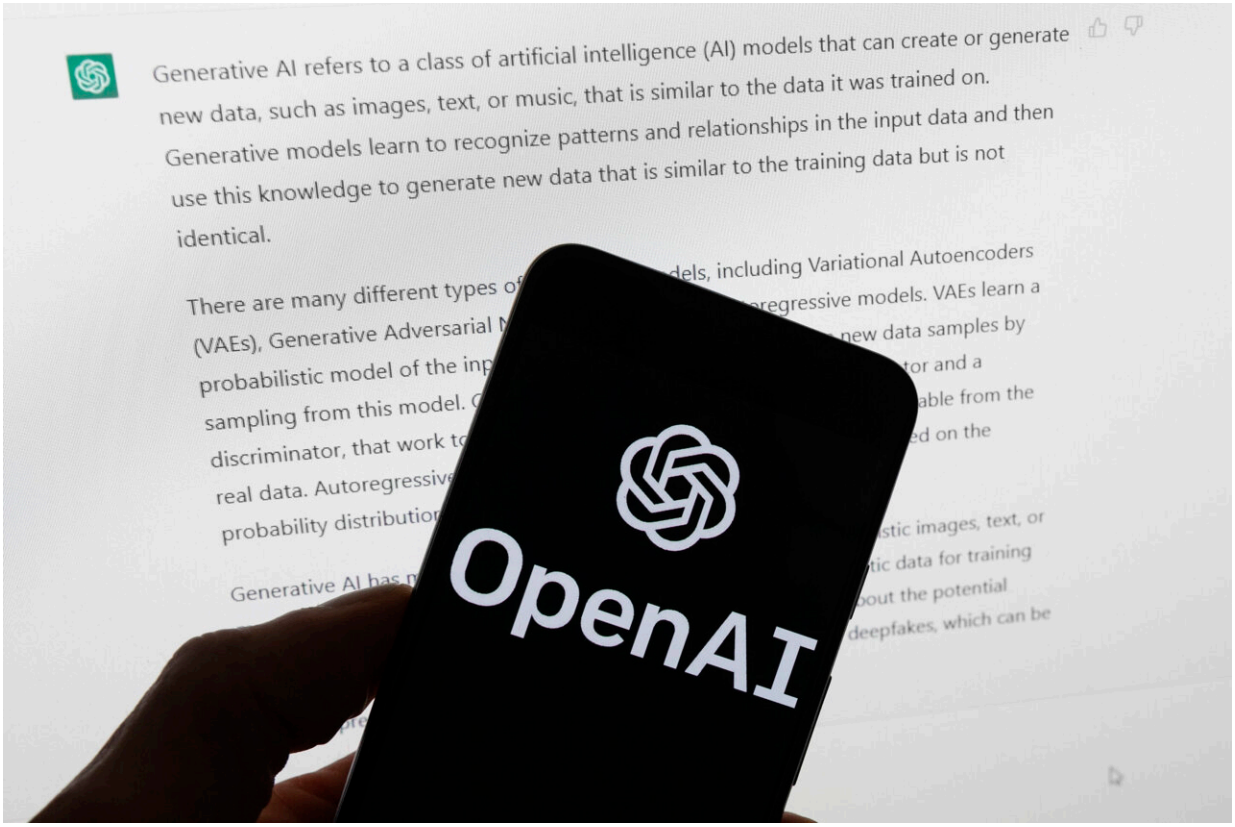
Jan Leike, who ran OpenAI's "Superalignment" team alongside a company co-founder who also resigned this week, wrote in a series of posts on the social media platform X that he joined the San Francisco-based company because he thought it would be the best place to do AI research.

"However, I have been disagreeing with OpenAI leadership about the company's core priorities for quite some time, until we finally reached a breaking point," wrote Leike, whose last day was Thursday.

An AI researcher by training, Leike said he believes there should be more focus on preparing for the next generation of AI models, including on things like safety and analyzing the societal impacts of such technologies. He said building "smarter-than-human machines is an inherently dangerous endeavor" and that the company "is shouldering an enormous responsibility on behalf of all of humanity."

"OpenAI must become a safety-first AGI company," wrote Leike, using the abbreviated version of artificial general intelligence, a futuristic vision of machines that are as broadly smart as humans or at least can do many things as well as people can.

Open AI CEO Sam Altman wrote in a reply to Leike's posts that he was "super appreciative" of Leike's contributions to the company was "very sad to see him leave."



The OpenAI logo is seen on a mobile phone in front of a computer screen displaying output from ChatGPT, March 21, 2023, in Boston. OpenAI has introduced a new artificial intelligence model. It says it works faster than previous versions and can reason across text, audio and video in real time. Credit: AP Photo/Michael Dwyer, File

Leike is "right we have a lot more to do; we are committed to doing it," Altman said, pledging to write a longer post on the subject in the coming days.

The company also confirmed Friday that it had disbanded Leike's Superalignment team, which was launched last year to focus on AI risks, and is integrating the team's members across its research efforts.

Leike's resignation came after OpenAI co-founder and chief scientist Ilya Sutskever said Tuesday that he was leaving the company after nearly a decade. Sutskever was one of four board members last fall who voted to push out Altman—only to quickly reinstate him. It was Sutskever who told Altman last November that he was being fired, but he later said he regretted doing so.

Sutskever said he is working on a new project that's meaningful to him without offering additional details. He will be replaced by Jakub Pachocki as chief scientist. Altman called Pachocki "also easily one of the greatest minds of our generation" and said he is "very confident he will lead us to make rapid and safe progress towards our mission of ensuring that AGI benefits everyone."

On Monday, OpenAI showed off the latest update to its artificial intelligence model, which can mimic human cadences in its verbal responses and can even try to detect people's moods.