OpenAI putting ‘shiny products’ above safety, says departing researcher

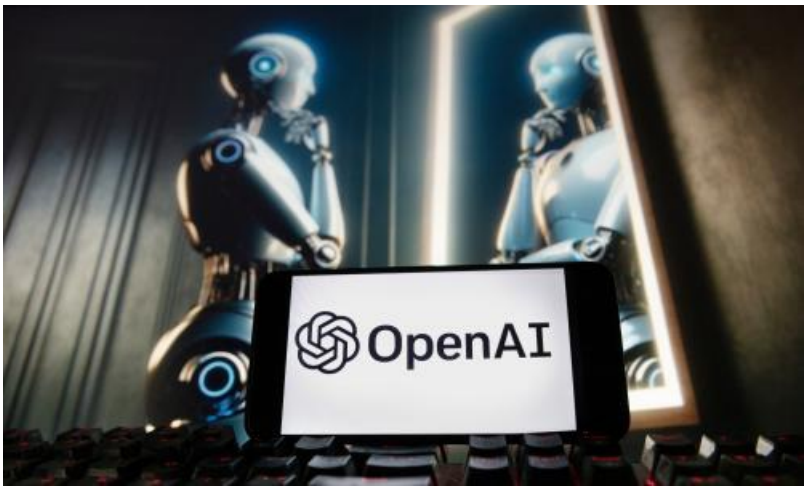
Publication Date: 2024-05-18

Author: Dan Milmo

Section: Technology

Tags: Artificial intelligence (AI), The Observer, ChatGPT, OpenAI, Computing, news

Article URL: https://www.theguardian.com/technology/article/2024/may/18/openai-putting-shiny-products-above-safety-says-departing-researcher



A former senior employee at OpenAI has said the company behind ChatGPT is prioritising “shiny products” over safety, revealing that he quit after a disagreement over key aims reached “breaking point”. Jan Leike was a key safety researcher at OpenAI as its co-head of superalignment, ensuring that powerful artificial intelligence systems adhered to human values and aims. His intervention comes before a global artificial intelligence summit in Seoul next week, where politicians, experts and tech executives will discuss oversight of the technology. Leike resigned days after the San Francisco-based company launched its latest AI model, GPT-4o. His departure means two senior safety figures at OpenAI have left this week following the resignation of Ilya Sutskever, OpenAI’s co-founder and fellow co-head of superalignment. Leike detailed the reasons for his departure in a thread on X posted on Friday, in which he said safety culture had become a lower priority. “Over the past years, safety culture and processes have taken a backseat to shiny products,” he wrote. OpenAI was founded with the goal of ensuring that artificial general intelligence, which it describes as “AI systems that are generally smarter than humans”, benefits all of humanity. In his X posts, Leike said he had been disagreeing with OpenAI’s leadership about the company’s priorities for some time but that standoff had “finally reached a breaking point”. Leike said OpenAI, which has also developed the Dall-E image generator and the Sora video generator, should be investing more resources on issues such as safety, social impact, confidentiality and security for its next generation of models. “These problems are quite hard to get right, and I am concerned we aren’t on a trajectory to get there,” he wrote, adding that it was getting “harder and harder” for his team to do its research. “Building smarter-than-human machines is an inherently dangerous endeavour. OpenAI is shouldering an enormous responsibility on behalf of all of humanity,” Leike wrote, adding that OpenAI “must become a safety-first AGI company”. Sam Altman, OpenAI’s chief executive, responded to Leike’s thread with a post on X thanking his former colleague for his contributions to the company’s safety culture. “He’s right we have a lot more to do; we are committed to doing it,” he wrote. Sutskever, who was also OpenAI’s chief scientist, wrote in his X post announcing his departure that he was confident OpenAI “will build AGI that is both safe and beneficial” under its current leadership. Sutskever had initially supported the removal of Altman as OpenAI’s boss last November, before backing his reinstatement after days of internal tumult at the company. Leike’s warning came as a panel of international AI experts released an inaugural report on AI safety, which said there was disagreement over the likelihood of powerful AI systems evading human control. However, it warned that regulators could be left trailing by rapid advances in the technology, warning of the “potential disparity between the pace of technological progress and the pace of a regulatory response”.