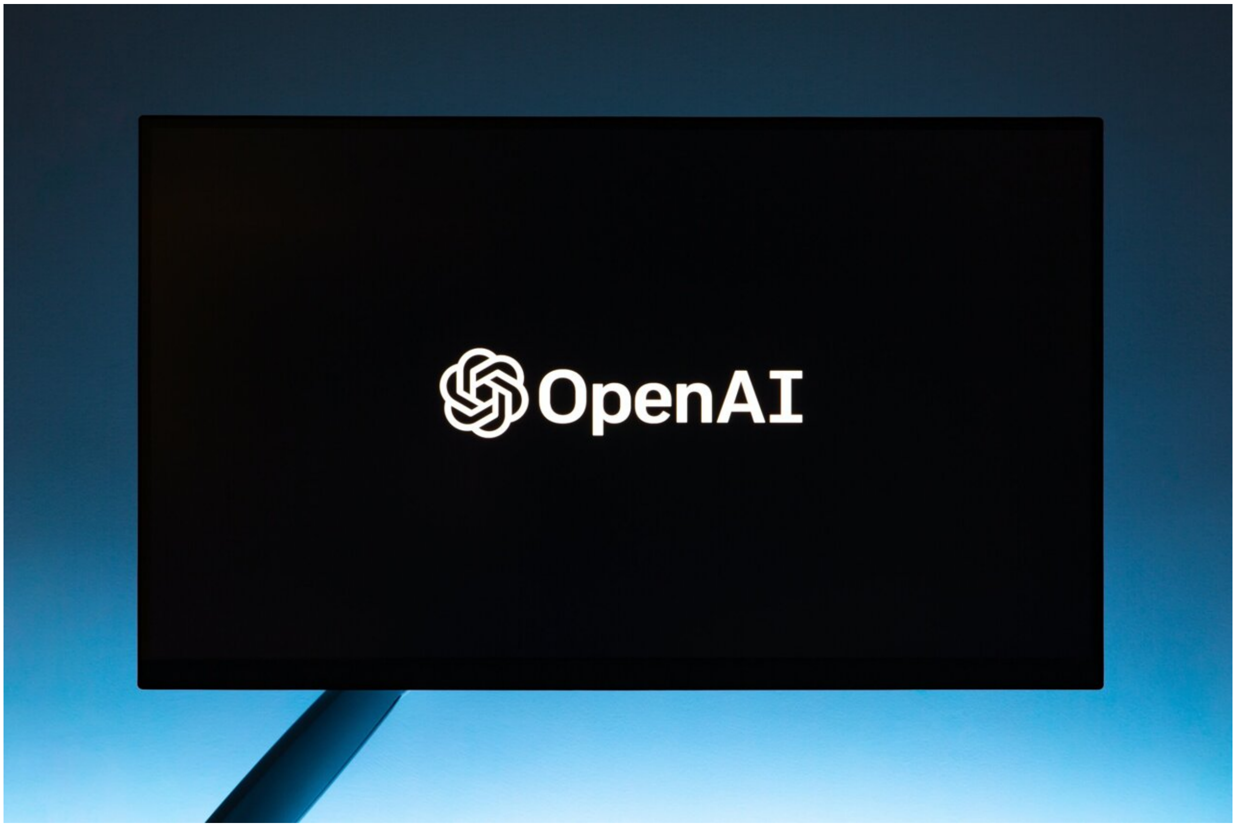
TechXplore

OpenAI says state-backed actors used its AI for disinfo

May 30 2024



Credit: Unsplash/CC0 Public Domain

OpenAI, the company behind ChatGPT, said Thursday it has disrupted five covert influence operations over the past three months that sought to use its artificial intelligence models for deceptive activities.

In a blog post, OpenAI said the disrupted campaigns originated from Russia, China, Iran, and a private company based in Israel.

The threat actors attempted to leverage OpenAI's powerful language models for tasks like generating comments, articles, social media profiles, and debugging code for bots and websites.

The company led by CEO Sam Altman said these operations "do not appear to have benefited from meaningfully increased audience engagement or reach as a result of our services."

Companies like OpenAI are under close scrutiny over fears that apps like ChatGPT or image generator Dall-E can generate deceptive content within seconds and in high volumes.

This is especially a concern with major elections about to take place across the globe and countries like Russia, China and Iran known to use covert social media campaigns to stoke tensions ahead of polling day.

One disrupted op, dubbed "Bad Grammar," was a previously unreported Russian campaign targeting Ukraine, Moldova, the Baltics and the United States.

It used OpenAI models and tools to create short political comments in Russian and English on Telegram.

The well-known Russian "Doppelganger" operation employed OpenAI's artificial intelligence to generate comments across platforms like X in languages including English, French, German, Italian and Polish.

OpenAI also took down the Chinese "Spamouflage" influence op which abused its models to research social media, generate multi-language text, and debug code for websites like the previously unreported revealscum.com.

An Iranian group, the "International Union of Virtual Media," was disrupted for using OpenAI to create articles, headlines and content posted on Iranian state-linked websites.

Additionally, OpenAI disrupted a commercial Israeli company called STOIC, which appeared to use its models to generate content across Instagram, Facebook, Twitter and affiliated websites.

This campaign was also flagged by Facebook-owner Meta earlier this week.

The operations posted across platforms like Twitter, Telegram, Facebook and Medium, "but none managed to engage a substantial audience," OpenAI said.

In its report, the company outlined AI leverage trends like generating high text/image volumes with fewer errors, mixing AI and traditional content, and faking engagement via AI replies.

OpenAI said collaboration, intelligence sharing and safeguards built into its models allowed the disruptions.

Citation: OpenAI says state-backed actors used its AI for disinfo (2024, May 30) retrieved 27 June 2024 from https://techxplore.com/news/2024-05-openai-state-actors-ai-disinfo.html