

OpenAI says it disrupted five covert influence operations that used its models

Source copy kept by wheresthe.ai for the story "OpenAI says it disrupted five covert influence operations that used its models" (<https://wheresthe.ai/d/openai-covert-influence-operations-report-2024-05/>).

Source: <https://openai.com/index/disrupting-deceptive-uses-of-AI-by-covert-influence-operations/>

Text saved 18 Oct 2026 (the page text through a reader, because the site refused a direct fetch).

SHA-256 of the saved text: e1fd3e44e5b8010b83a274fc0a9b5cfccc24dd40baebc874e79648873c47832c

OpenAI is committed to enforcing policies that prevent abuse and to improving transparency around AI-generated content. That is especially true with respect to detecting and disrupting covert influence operations (IO), which attempt to manipulate public opinion or influence political outcomes without revealing the true identity or intentions of the actors behind them.

In the last three months, we have disrupted five covert IO that sought to use our models in support of deceptive activity across the internet. As of May 2024, these campaigns do not appear to have meaningfully increased their audience engagement or reach as a result of our services.

This blog describes the threat actors we disrupted, attacker trends we identified, and important defensive trends - including how designing AI models with safety in mind in many cases prevented the threat actors from generating the content they desired, and how AI tools have made our own investigations more efficient. Alongside this blog, we are publishing a trend analysis that describes the behavior of these malicious actors in detail.

[Read the full report](#)(opens in a new window)

Threat actors work across the internet. So do we. By collaborating with industry, civil society, and government we tackle the creation, distribution, and impact of IO content. Our investigations and disruptions were made possible in part because there's been so much detailed threat reporting over the years by distribution platforms and the open-source community. OpenAI is publishing these findings, as other tech companies do, to promote information sharing and best practices amongst the broader community of stakeholders.

Over the last three months, our work against IO actors has disrupted covert influence operations that sought to use AI models for a range of tasks, such as generating short comments and longer articles in a range of languages, making up names and bios for social media accounts, conducting open-source research, debugging simple code, and translating and proofreading texts.

Specifically, we disrupted:

* A previously unreported operation from Russia, which we dubbed **Bad Grammar**, operating mainly on Telegram and targeting Ukraine, Moldova, the Baltic States and the United States. The people behind Bad Grammar used our models to debug code for running a Telegram bot and to create short, political comments in Russian and English that were then posted on Telegram.

* An operation originating in Russia known as **Doppelganger**(opens in a new window). People acting on behalf of **Doppelganger** used our models to generate comments in English, French, German, Italian and Polish that were posted on X and 9GAG; translate and edit articles in English and French that were posted on websites linked to this operation; generate headlines; and convert news articles into Facebook posts.

* A Chinese network known as **Spamouflage**(opens in a new window), which used our models to research public social media activity, generate texts in languages including Chinese, English, Japanese and Korean that were then posted across platforms including X, Medium and Blogspot, and debug code for managing databases and websites, including a previously unreported domain, [revealscum\[.\]com](https://revealscum[.]com).

* An Iranian operation known as the **International Union of Virtual Media**(opens in a new window) **(IUVM)**, which used our models to generate and translate long-form articles, headlines and website tags that were then published on a website linked to this Iranian threat actor, [iuvmpress\[.\]co](https://iuvmpress[.]co);

* Activity by a commercial company in Israel called **STOIC**, because technically we disrupted the activity, not the company. We nicknamed this operation **Zero Zeno**, for the founder of the stoic school of philosophy. The

people behind Zero Zeno used our models to generate articles and comments that were then posted across multiple platforms, notably Instagram, Facebook, X, and websites associated with this operation.

The content posted by these various operations focused on a wide range of issues, including Russia's invasion of Ukraine, the conflict in Gaza, the Indian elections, politics in Europe and the United States, and criticisms of the Chinese government by Chinese dissidents and foreign governments.

So far, these operations do not appear to have benefited from meaningfully increased audience engagement or reach as a result of our services. Using Brookings' Breakout Scale,[\(opens in a new window\)](#) which assesses the impact of covert IO on a scale from 1 (lowest) to 6 (highest), none of the five operations included in our case studies scored higher than a 2 (activity on multiple platforms, but no breakout into authentic communities).