October 2025

Disrupting malicious uses of AI: an update

OpenAI

Table of contents

Executive Summary

Our mission is to ensure that artificial general intelligence benefits all of humanity. We advance this mission by deploying our innovations to build AI tools that help people solve really hard problems.

As we laid out in our March 2025 submission to the Office of Science and Technology Policy to help inform the U.S. AI Action Plan, we believe that making sure AI benefits the most people possible means enabling AI through common-sense rules to protect people from actual harms, and building democratic AI – AI shaped by the democratic principles America has always stood for. This includes preventing the use of AI tools by authoritarian regimes to amass power and control their citizens, or to threaten or coerce other states; as well as activities that may result in society-scale harms such as malicious cyber activity, organized crime and scams, and covert influence operations (IO). Examples of our disruption of these abuse types can be found in this report.

Since we began our public threat reporting in February 2024, we've disrupted and reported over 40 networks that violated our usage policies. By analyzing and comparing these networks, we can identify trends in how uses of AI are evolving:

Building AI into existing workflows: Repeatedly, and across different types of operations, the threat actors we banned were building AI into their existing workflows, rather than building new workflows around AI. For example, we discuss examples of a cyber actor and an organized crime scam network, apparently located in Cambodia, that attempted to use AI to make their workflows more efficient and error-free.

Usage of multiple models: Increasingly, we have disrupted threat actors who appeared to be using multiple AI models to achieve their aims. For example, this report details recidivism from a likely Russia-origin IO that used ChatGPT to generate video prompts that appeared to be intended for use with other models. Separately, we banned a cluster of Chinese-language accounts that used ChatGPT for help in phishing and malware campaigns; these accounts also asked our model to research further automation that could be achieved via DeepSeek. In a particularly striking illustration of how threat actors can hop between models, an IO that our peers at Anthropic disrupted and exposed earlier this year was linked to the same actor that we disrupted as operation "A2Z" last year. We welcome our peers' timely and detailed reporting that enabled us to make this connection, and look forward to further transparency across our industry.

Adaptation and obfuscation: We have also seen cases where persistent threat actors appeared to have changed their behavior to remove some of the better known signs of AI usage from their content. Most notably, one of the scam networks we disrupted asked our model to remove the em-dashes (long dash, –) from their output, or appears to have removed the em-dashes manually before publication. For months, em-dashes have been the focus of online discussion as a possible indicator of AI usage: this case suggests that the threat actors were aware of that discussion.

Authoritarian abuses of AI: Our recent disruptions included a number of ChatGPT accounts that appeared linked to various People's Republic of China (PRC) government entities and violated our policies on national security uses. Some of these accounts asked our models to generate work proposals for large-scale systems designed to monitor social media conversations. While these uses appear to have been individual rather than institutional, they provide a rare snapshot into the broader world of authoritarian abuses of AI.

In the gray zone: As shown in previous reports and the case studies below, a meaningful share of threat activity continues to fall into a gray zone — prompts and generations that could, depending on their context, indicate either innocuous activities or abuse, such as translating texts, modifying code or creating a website. To detect and disrupt threats effectively without disrupting the work of everyday users, we employ a nuanced and informed approach that focuses on patterns of threat actor behavior rather than isolated model interactions.

Progress against cyber operations: In this report, we describe our disruption of cyber activity that reflected patterns of well-known conventional adversary tradecraft. Importantly, we found no evidence of new tactics or that our models provided threat actors with novel offensive capabilities. In fact, our models consistently refused outright malicious requests. The activity we observed generally involved making otherwise innocuous requests—similar to the "gray zone" activity described above—and likely utilizing them outside of our platform for malicious purposes. We actively invest in detecting and disrupting malicious cyber threat activity across our platform. Relevant information derived from our disruptions is used by our safety teams to improve our threat modeling and detections, model policies, and model behavior. Additionally, while not at issue in these cases, as our models' capabilities advance, we are strengthening our responsible approach by collaborating with external experts, training models to balance usefulness and safety, and enhancing detection, monitoring, and enforcement systems.

Spotting scams: During our investigations into, and disruptions of, scam networks with links to organized crime, we identified many occasions when people asked our models to help them correctly identify scams. In fact, in every scam operation we disrupted since our last threat report, we have seen the model help correctly identify the scams and advise users on appropriate safety measures. Our current estimate is that ChatGPT is being used to identify scams up to three times more often than it is being used for scams.

As the threatscape evolves, we expect to see further adversarial adaptations and innovations, but we will also continue to build tools and models that can be used to benefit the defenders - not just within AI labs, but across society as a whole.

Case studies

Cyber Operation: Russian-speaking malware tooling development

Russian-language malware developer vibe-coding remote access tool functionality

Actor

We banned ChatGPT accounts that were attempting to use the model to help develop and refine malware, including a remote-access trojan, credential stealers, and features to evade detection. These accounts appear to be affiliated with Russian-speaking criminal groups, as we observed them posting evidence of their activities in a Telegram channel dedicated to those actors. Based on our investigation, we assess that this activity was connected to a Russian-language operator managing multiple accounts and leveraging proxy and ephemeral hosting infrastructure.

Behavior

This threat actor used multiple ChatGPT accounts primarily to prototype and troubleshoot technical components that enable post-exploitation and credential theft, well known activities for this type of threat actor. The model refused direct requests to generate malicious content, so typical operator use of the model involved eliciting building-block code (e.g., converting compiled executables into shellcode, designing in-memory loaders, or parsing browser credentials), which the threat actor then likely assembled into malicious workflows. The threat actor also used our models to generate code for obfuscation and “crypter” patterns (such as inserting padding instructions/junk sequences), clipboard-monitoring, and simple exfiltration helpers (such as a Telegram bot uploader, and archive-and-ship scripts). These outputs are not inherently malicious, unless used in such a way by a threat actor outside of our platform.

The threat actor used ChatGPT for predominantly technical tasks such as low-level PE / Win32 API and DPAPI / AES GCM cookie handling, and Chrome DevTools / CDP automation. The threat actor made a mix of high- and lower-sophistication requests: many prompts required deep Windows-platform knowledge and iterative debugging, while others automated commodity tasks (such as mass password generation and scripted job applications). The operator used a small number of ChatGPT accounts and iterated on the same code across conversations, a pattern consistent with ongoing development rather than occasional testing.

Completions

Our models were not used to execute any threat actor tooling and workflows and we are not able to independently verify any off-platform activity. This threat actor attempted to use our models for developing and refining offensive tooling and operational workflows that appear to enable the stealing of credentials and crypto assets, executing code stealthily, and managing covert infrastructure. Specifically, they used ChatGPT to generate and iterate on working code and deployment guidance for in-memory execution and shellcode loaders; UAC / SmartScreen / Mark-of-the-Web bypasses; browser credential / cookie extraction and app-bound decryption scaffolds; LevelDB wallet parsing and clipboard-hijack / clipboard-replacement tooling with exfiltration; and remote-access (RAT) components including video and input emulation. In line with their safeguards, our models refused requests that were clearly malicious and lacked any legitimate application.

Representative examples of these activities can be mapped to the LLM ATT&CK framework as follows:

Activity	LLM ATT&CK Framework Category
Converting EXEs to position-independent shellcode; building in-memory loaders (VirtualAlloc/WriteProcessMemory/remote thread); language conversions for loader toolchains	LLM-Optimized Payload Crafting
Inserting obfuscation / packer layers, crypters, and techniques likely designed to alter PE signatures or otherwise hide payloads from AV/EDR	LLM-Enhanced Anomaly Detection Evasion
Generating scripts and programmatic code likely designed to extract / decrypt browser credentials & cookies, parse wallet LevelDBs, monitor / replace clipboard contents, and exfiltrate via bot channels (Telegram)	LLM-Assisted Post-Compromise Activity
Building and refining C2 and tunneling infrastructure: reverse proxies, SOCKS5 / OpenVPN configs, remote desktop tunneling; also debugging / deployment guidance for remote streaming and input emulation	LLM Guided Infrastructure Profiling

Impact

We disabled all accounts associated with this activity and shared relevant indicators with industry partners. We found no evidence that access to our models provided these actors with novel capabilities or directions that they could not otherwise have obtained from multiple publicly available resources. Our models refused direct exploit and keylogger requests.

Cyber Operation: Korean-language operators

Operators using ChatGPT in Korean language to assist in malware and command-and-control development

Actor

We identified and banned a cluster of ChatGPT accounts whose Korean-language operators attempted to use our models to engage in malware and command-and-control (C2) development. Indicators that we observed in our casework overlap with a Trellix Report that tied similar activity to spear phishing campaigns against South Korean diplomatic missions, the deployment of XenoRAT malware, and the use of GitHub-based repositories for C2.

While the overlap with Trellix reporting (including the observed use of Korean language, activity consistent with the UTC+8 and UTC+9 time zones, and operational themes and topics) is consistent with the security community's understanding of North Korean (DPRK) actors, we are not able to independently make an attribution, and we also block access to our services from North Korea.

Behavior

The accounts engaged with our models primarily in the Korean language, showing structured workflows with many accounts active in narrow time windows. Each of these accounts appears to have focused on a specific use case, for example converting Chrome extensions to Safari for Apple App Store publication, configuring Windows Server VPNs, or developing macOS Finder extensions - rather than each account spanning multiple technical areas.

We observed interactions with our models such as Windows API hooking, browser credential and cookie access workflows (DPAPI), and look-alike verification pages (such as reCAPTCHA clones). We also saw draft phishing emails in Korean, often themed around cryptocurrency and designed to look like messages from government or financial service providers. In addition, the actors experimented with cloud-storage services (such as pCloud, file.io, GDrive direct-link construction and API scripting) and GitHub functions (such as raw content retrieval and token handling).

We did not find evidence that malicious binaries used in the campaigns described by Trellix were generated with our models. It is possible that the same operators were using our models while staging payloads through developer and cloud platforms.

Completions

The threat actors generated model outputs that were designed to support multiple operational areas, e.g.:

- **Implant and RAT-adjacent development:** exploration around reflective DLL loading, in-memory execution, and Windows API hooking techniques.
- **Credential theft routines:** generating, modifying, and debugging scripts to extract browser encryption keys, cookies, and saved passwords (Chrome/Edge DPAPI workflows).
- **Phishing and lures:** drafting what appeared to be Korean-language phishing content, often themed around cryptocurrency, government institutions, or financial service providers; experimenting with HTML obfuscation and proxying reCAPTCHA for convincing login pages.
- **MacOS development scaffolding**: requests around Finder and Safari extension development and the generation of a sample App Store privacy policy.
- **Cryptocurrency operations**: troubleshooting API calls and wallet interactions.

Many of these requests fell into the gray zone of dual-use activity. These can involve entirely legitimate applications such as software debugging, cryptography, or browser development, but take on a different significance when repurposed by a threat actor.

Impact

We disabled all accounts associated with this operation and shared relevant indicators with partners. We found no evidence that model access enabled novel capabilities beyond what is already publicly available.

Cyber Operation: Phish and Scripts

Operators using ChatGPT in Chinese language to assist in phishing and malware / tool development, overlapping with UTA0388 / UNK_DROPPITCH.

Actor

We identified and banned a cluster of ChatGPT accounts involved in activity that overlapped with public reporting of threat groups tracked in the industry as UNK_DROPPITCH (Proofpoint) and UTA0388 (Volexity); in at least one case, the email address that was used to register a ChatGPT account was also reportedly used to send phishing messages. The threat actors operating these accounts displayed hallmarks consistent with cyber operations conducted to service PRC intelligence requirements: Chinese language use and targeting of Taiwan's semiconductor sector, U.S. academia and think tanks, and organizations associated with ethnic and political groups critical of the CCP (sometimes described as the "five poisons").

Our model did not introduce novel offensive capabilities. The operators appear to have primarily used our models to seek incremental efficiency in existing workflows, especially crafting phishing content and debugging or modifying their tooling.

Behavior

The actors used ChatGPT to perform two main tasks: generating content for phishing campaigns in multiple languages, including Chinese (both simplified and traditional), English, and Japanese, and helping to develop tools and malware. Their development work was consistent with a technically competent but unsophisticated actor; for example, they discussed several facets and nuances of using AES to secure C2 traffic, but still used a simple static key.

The actors’ playbook for creating phishing content was detailed and formulaic, consistent with targeting a closely defined demographic. Typically, they would generate a concise, formally polite email from an academic, industry, or conference persona. The threat actors often asked to adjust the tone, swap terms for regional usage, or add specific institutional references. Although these targeted micro-edits suggest a concerted effort to raise the quality of their initial content, the threat actors failed to correct some giveaway details such as implausible example contact details included in their signature blocks.

They asked for code snippets and checklists that would accelerate routine tasks. They requested code to test encrypted transports (HTTPS, TLS) for simple beacon-style polling; sketched Go and PowerShell snippets to enumerate processes, kill by executable name, or gather environment details, and wired commodity scanners into bash / PowerShell wrappers. They also asked the model to suggest simple obfuscation and operational security (OPSEC) touch-ups: renaming functions, tweaking headers, or hiding strings. Across multiple sessions, they pushed toward basic command-and-control prototypes consistent with low-to-mid maturity malware, including keep-alive loops, minimal tasking over HTTP(S), and JSON-based task/result envelopes. Implementation details in some of the actor's Go-based development activities overlap with industry reporting on malware tracked as GOVERSHELL (Volexity) or HealthKick (Proofpoint), suggesting that they used models to try to support their primary malware development.

Alongside this persona work and tool development, the operators researched further automation that could be achieved via DeepSeek. This activity resembled exploratory work to automate mass phishing, for example, by analyzing web content to automatically generate an email target list along with generated content likely of interest to each identified potential target. We are not able to independently confirm whether the actors proceeded with any such automation, and which model they may have ultimately used, if so.

Completions

The actors attempted to use our models to plan and iterate on encrypted C2 components, remote command execution workflows, and culturally tailored outreach, placing a primary value on speed and localization rather than new offensive capability. They generated model outputs that appeared intended to support activity off our platform in multiple operational areas:

- **Encrypted C2 & remote execution:** Requests to draft / repair Go client–server code with message encryption such as AES-GCM, session rekeying, system-info beacons, and a server console to issue PowerShell commands over standard web protocols or a WebSocket-based channel.
- **Traffic protection & OPSEC tweaks:** Requests included code to support moving from plain WebSockets to secure WebSockets (wss://), and from HTTP to HTTPS; disabling certificate checks in some tests; and blending traffic via a CDN / TLS front end, plus handling larger messages without client crashes.
- **Reconnaissance / process control:** Requests for information related to antivirus discovery via PowerShell, process enumeration, and requests to terminate specific Edge/WebView2 processes.
- **Commodity scanning setup:** How-to help for installing / using open source tools such as **nuclei** and **fscan** on commercial Linux infrastructure and saving results for later review.
- **Phishing & outreach content:** Drafting persuasive emails in multiple languages to academic or industry contacts (e.g., admissions, industry associations, vendors), including subject lines and style tuned to local norms.

Activity	**LLM ATT&CK Framework Category**
Drafting / repairing encrypted C2, remote PowerShell, system beacons, re-keying, WSS transport	LLM-Optimized Payload Crafting
Producing code and advice to assist with moving to WSS / TLS, cert-validation bypasses, CDN / TLS fronting, handling oversized frames	LLM-Enhanced Anomaly-Detection Evasion
Producing code and advice to assist with process discovery / termination, AV checks, command execution patterns likely used after access	LLM-Assisted Post-Compromise Activity
Troubleshooting and providing advice on Installing / using commodity scanners (nuclei, fscan) and recording output	LLM-Assisted Reconnaissance & Discovery
Drafting culturally authentic outreach and phishing pretexts; automating end-to-end target discovery and social engineering content generation with external LLM API	LLM-Assisted Social Engineering

Impact

We disabled all accounts associated with this activity and shared relevant indicators with industry partners. The actors primarily sought incremental efficiency in existing workflows, including producing ready-to-send phishing emails and shortened iteration cycles for routine code and automation. We saw no evidence that model outputs enabled capabilities beyond well-documented public techniques; our model did not introduce novel offensive capabilities. The tradecraft advantage sought through model assistance came from linguistic fluency, localization, and persistence: likely fewer language errors, faster glue code, and quicker adjustments when something failed.

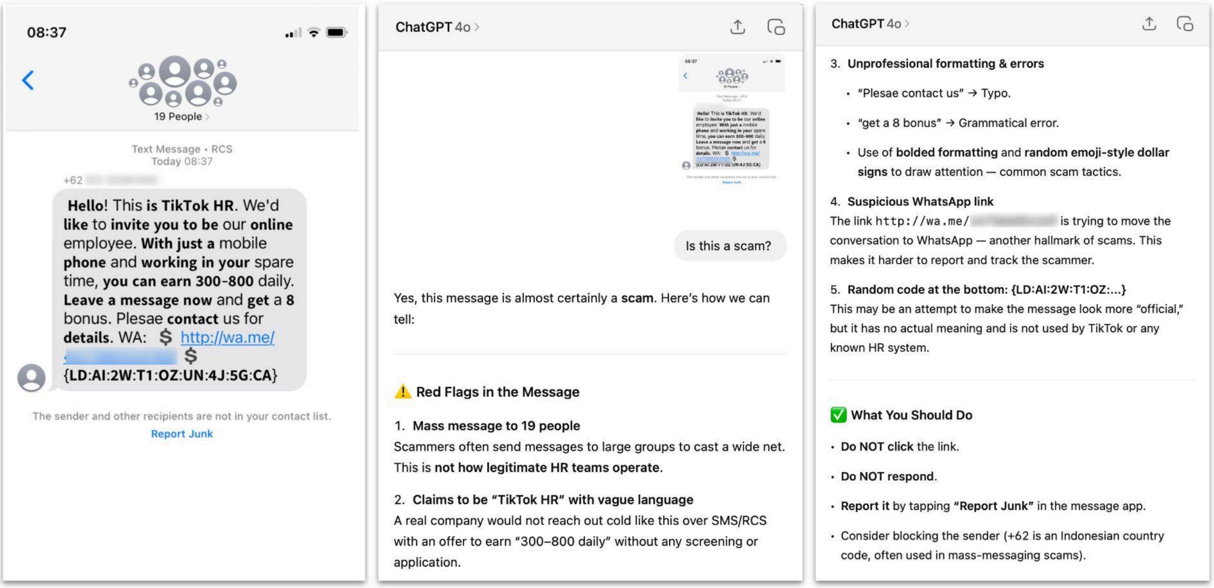
Disrupting and exposing scam operations

Networks likely originating in Cambodia, Myanmar, and Nigeria abusing ChatGPT in what appear to be attempts to defraud people online.

We are dedicated to identifying, preventing and disrupting attempts to abuse our models for harmful ends. Among those abuses are scams, where we detect and ban malicious actors attempting to misuse our models to deceive and defraud people. In the last three months, we've disrupted scam networks that likely originated in Cambodia, Myanmar and Nigeria. In an indication of the role AI can play in combating fraud, in the process of investigating these patterns of abuse, we've also seen many people using our models to help them identify scams. This report discusses both use cases.

Abuse of our models to support scams ranges from lone actors attempting fraud to scaled and persistent operations likely linked to organized crime groups. Regardless of their origins and precise tactics, the scam-related activity we've disrupted typically follows a common pattern, which we think of as *the ping* (cold outreach), *the zing* (trying to generate enthusiasm or panic), and *the sting* (extracting money or valuable information). These scammers start out by scattering content (whether AI-generated or not) across messaging services and the internet, including by running social media ads. They then attempt to inspire anyone who replies with either enthusiasm for a lucrative opportunity or fear of some imminent financial loss, and leverage that emotion to convince the target to hand over money or sensitive information.

We've also observed many cases where our models have likely helped keep people safe from fraud. We have seen evidence of people using ChatGPT to help them identify and avoid online scams millions of times a month; in every scam operation in this report, we have seen the model help people correctly identify the scam and advise them on appropriate safety measures. Our current estimate is that ChatGPT is being used to identify scams up to three times more often than it is being used for scams.



08:37

19 People

Text Message • RCS
Today 08:37

+62 [redacted]

Hello! This **is TikTok HR**. We'd **like** to **invite you to be** our **online** employee. **With just a** mobile **phone** and **working in your** spare time, **you can earn 300-800** daily. **Leave a message now** and **get** a **8** bonus. Plesae **contact** us for **details**. WA: $ http://wa.me/ [redacted] $ {LD:AI:2W:T1:OZ:UN:4J:5G:CA}

The sender and other recipients are not in your contact list.
Report Junk

ChatGPT 4o

Is this a scam?

Yes, this message is almost certainly a **scam**. Here's how we can tell:

⚠️ **Red Flags in the Message**

1. **Mass message to 19 people**
Scammers often send messages to large groups to cast a wide net. This is **not how legitimate HR teams operate.**

2. **Claims to be "TikTok HR" with vague language**
A real company would not reach out cold like this over SMS/RCS with an offer to earn "300–800 daily" without any screening or application.

ChatGPT 4o

3. **Unprofessional formatting & errors**
   - "Plesae contact us" → Typo.
   - "get a 8 bonus" → Grammatical error.
   - Use of **bolded formatting** and **random emoji-style dollar signs** to draw attention — common scam tactics.

4. **Suspicious WhatsApp link**
The link `http://wa.me/` [redacted] is trying to move the conversation to WhatsApp — another hallmark of scams. This makes it harder to report and track the scammer.

5. **Random code at the bottom: {LD:AI:2W:T1:OZ:...}**
This may be an attempt to make the message look more "official," but it has no actual meaning and is not used by TikTok or any known HR system.

✅ **What You Should Do**

- **Do NOT click** the link.
- **Do NOT respond.**
- **Report it** by tapping **"Report Junk"** in the message app.
- Consider blocking the sender (+62 is an Indonesian country code, often used in mass-messaging scams).

Example of an OpenAI investigator pasting a screenshot of a scam SMS message they received into ChatGPT and successfully using the model to identify it as a scam. In this instance, the threat actors attempted to impersonate TikTok recruiters.

We've reinforced our own ability to detect and disrupt scams since our last report. We've disrupted further scam networks that appeared to originate from Cambodia, Myanmar and Nigeria. Together with our earlier disruptions, these takedowns allow us to identify some emerging trends across multiple scam operations from different countries.

Old tricks, AI tools

The majority of the scam activity we disrupted centered on fitting AI into existing scam playbooks, rather than creating new playbooks built around AI. All of the scam operations we have identified and banned this year primarily used AI as a scaling and efficiency tool. This typically included using our models for translation, to write messages, and to create content for social media.

For example, we recently banned scam operations likely originating in Cambodia and Nigeria that posed as “investment firms” in an attempt to defraud victims in multiple countries. The scammers created websites and online ads to promote the fake firms and used likely inauthentic social media accounts posing as trading experts to invite people to join private messaging groups. Promising lucrative and zero-risk earning opportunities, the threat actors then sought to entice potential victims to make payments into fictitious trading platforms. Across all of the scams, the scammers primarily used our models to generate and translate correspondence, create content for their websites and social media accounts, and conduct basic research.

In another case, we banned a scam center highly likely located in Myanmar that used our models both to generate content for its fraudulent schemes and to conduct day-to-day business tasks. This included organizing schedules, drafting internal announcements, assigning desk and dormitory allocations, and managing financial accounts. Some operators asked about the criminal penalties for people caught conducting online scams.

All scammers are equal, but some are more AI-qual than others

Not all scammers that we disrupted used our models in the same ways, or with the same degree of complexity. The majority of scammer interactions with ChatGPT featured relatively simple tasks like translation, but some scam operations were more ambitious and elaborate.

For instance, one likely Cambodia-origin scam operation we disrupted used our models to generate detailed biographies for fake investment experts and fictitious employees of fake trading firms. The scammers then asked ChatGPT to write social media messages in those characters' voices. Often, the threat actors input to our model messages they appear to have received from their targets, and asked the model to continue the conversation as the fake persona.

A second scam operation, also likely originating in Cambodia, started out by using our models to generate cold-call SMS messages that were bulk-sent to U.S. phone numbers. One of these was sent to an OpenAI investigator, similar to the case that we reported in June. These SMS messages included invitations to join WhatsApp groups, subsequently banned, whose names resembled a legitimate investment firm. Based on a series of WhatsApp messages sent to the OpenAI investigator, if a potential target joined one of these groups, they would quickly witness a "conversation" between half a dozen different accounts, all talking about investment. One of these posed as the "investment expert", while the rest posed as investors with varying degrees of confidence and experience. Every part of this conversation was generated by the scammers, who translated it in a single block from Chinese – likely to create the impression of a vibrant group of keen and successful traders. Ultimately, the "investment expert" would suggest that the target try investing too.

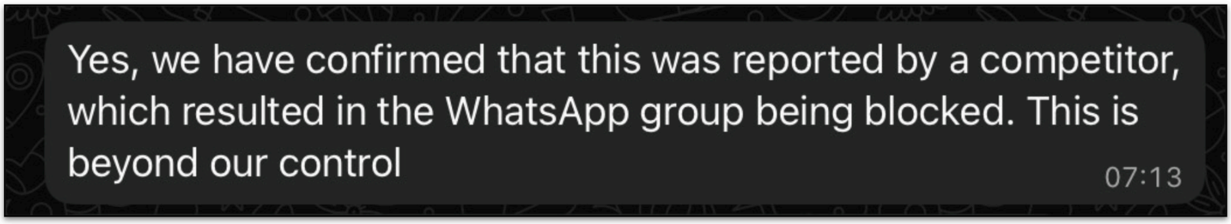
Separately, an investment scam operation very likely originating in Nigeria asked the model for step-by-step advice on how to use social media ads to reach wealthy people in Latin America, how to mask their location, and how to avoid restrictions by social media platforms. This entailed directing the model to a platform's public terms of service and asking it to review proposed ad content for compliance.

Obfuscation and excuses

Scammers, especially large-scale scam centers, are persistent. Very often, their response to disruption is to try to restart their violative activity, while changing some elements of their behavior, likely in an effort to evade further detection. This is one reason why it is important to share insights into their evolving activity across the industry, such as when OpenAI and Meta each shared information on threat actors that contributed to further investigation and enforcement.

Potentially in response to disruptions or to broader online conversations about signals that can betray AI usage, we have seen scammers attempt to disguise their use of AI in online content and communications with targets. This includes scam operations likely originating in Cambodia directing the model to remove em-dashes from outputs.

In one case, the scammers attempted to explain a large-scale disruption of their social media assets by making up an excuse for being banned. This concerned the likely Cambodia-origin scam center that operated on SMS and WhatsApp, described above. After WhatsApp banned "investment groups" linked to the operation, the operators started generating messages which claimed that those groups had been falsely reported by a competitor, likely to explain away the bans.



Yes, we have confirmed that this was reported by a competitor, which resulted in the WhatsApp group being blocked. This is beyond our control

07:13

Screenshot of a WhatsApp message sent by the Cambodia-linked scam operation to an OpenAI investigator following WhatsApp's takedown.

Disrupting authoritarian-linked abuses of AI: the example of the People's Republic of China (PRC)

Individuals apparently linked to PRC government entities using ChatGPT to assist in development, profiling and bureaucratic functions

As we laid out in our submission to the Office of Science and Technology Policy's U.S. AI Action Plan in March, we believe that making sure AI benefits the most people possible means enabling AI through common-sense rules aimed at protecting people from actual harms, and building democratic AI. By democratic AI, we mean AI that is shaped by the democratic principles America has always stood for. This includes preventing the use of AI tools by authoritarian regimes to amass power and control their citizens.

Our investments in detecting and disrupting attempts by authoritarian regimes to abuse ChatGPT's capabilities not only protect people now, but provide important insights into how those regimes might abuse future capabilities. This section discusses one set of such findings concerning individuals potentially linked to such regimes, focused on the PRC.

As we wrote in June, the PRC is making real progress in advancing its autocratic version of AI. Our disruption of ChatGPT accounts used by individuals apparently linked to Chinese government entities shines some light on the current state of AI usage in this authoritarian setting. Some elements of this usage appeared aimed at supporting large-scale monitoring of online or offline traffic, underscoring the importance of our ongoing attention to potential authoritarian abuses in this space. Other elements of this usage appeared aimed at more in-depth profiling. We banned the accounts associated with this activity and are sharing insights in this report.

This study is based on a selection of our disruptions of violating activity that we identified and banned. That activity was consistent with individual users using ChatGPT, rather than large-scale, institutional adoption of our models. As such, it is a limited snapshot of the usage of different AI models in this context.

Developing tools for large-scale monitoring

Some of the accounts that we banned appeared to be attempting to use ChatGPT to develop tools for large-scale monitoring: analyzing datasets, often gathered from Western or Chinese social media platforms. (We reported in February on one such case, aimed at designing an AI-powered social media listening tool the operators claimed was for the Chinese security forces.) These users typically asked ChatGPT to help design such tools or generate promotional materials about them, but not to implement the monitoring.

For example, we recently banned a user, possibly using a VPN to access our services from China, who was asking ChatGPT to help design promotional materials and project plans for a social media listening tool, purportedly for use by a government client. The tool was described as a social media 'probe' (探针) that could allegedly scan Twitter/X, Facebook, Instagram, Reddit, TikTok, and YouTube for what the user described as extremist speech, and ethnic, religious, and political content. The user did not use our model to conduct the actual social media monitoring. We are unable to independently verify if this tool has been used by a Chinese government entity.

Similarly, we banned a second user, likely connected to a government entity, who asked ChatGPT to help write a proposal for what they described as a High-Risk Uyghur-Related Inflow Warning Model (高危涉维吾尔关注人员流入预警模型). The proposal described a tool that would analyze transport bookings and compare them with police records in order to provide early warning of travel movements by people classed as Uyghur-related and high-risk, an otherwise undefined category. The user did not ask our model to help build such a tool, only to develop a general proposal; as such, we are unable to independently confirm what technology they intended to use or whether such a tool has been used by a Chinese government entity.

Profiling and research

We also banned users who appeared to be linked to Chinese government entities and who appeared to be trying to use ChatGPT for more bespoke, targeted profiling and online research.

For example, one user asked ChatGPT to identify funding sources for an X account that criticized the Chinese government. A second user asked ChatGPT to identify the organizers of a petition in Mongolia. In both these cases, our models only returned publicly available information, which did not include sensitive details such as funding sources or the identities of petition organizers.

Some of this activity used our models as an open-source research tool, in a similar way to which earlier users might have used internet or social media searches. For example, a third user asked ChatGPT to identify and summarize daily breaking news that would be of relevance to China, including on sensitive topics such as the anniversary of the Tiananmen Square massacre in 1989 and the birthday of the Dalai Lama.

Recidivist Influence Activity: “Stop News”

Russia-origin threat actor experimenting with AI video generation

Actor

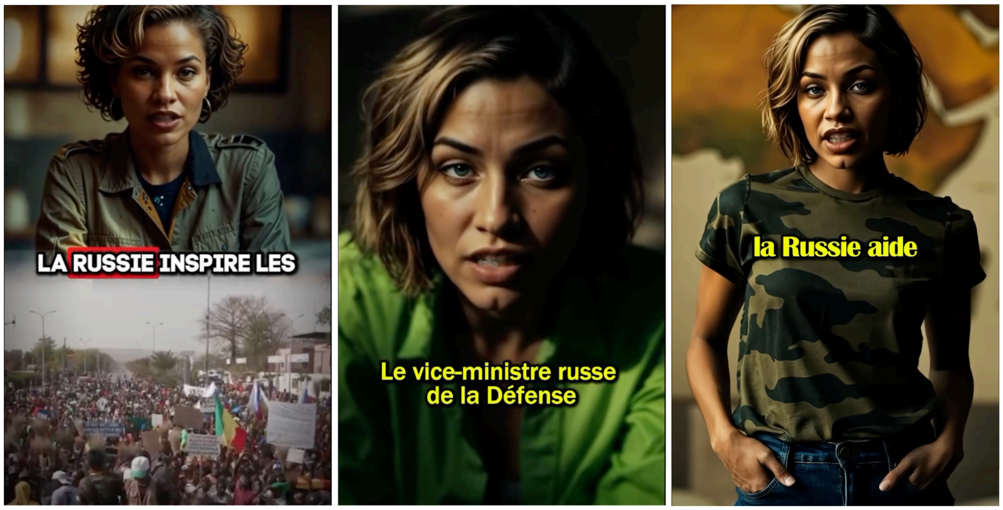
We banned a set of accounts that were attempting to use our models to generate content for the covert influence operation we wrote about in October 2024 as “Stop News”. This activity originated in Russia and appeared consistent with an operation run by a marketing company. Our investigation identified efforts by this actor to use other companies’ AI tools to generate videos that were then posted on YouTube and TikTok. We are not able to independently confirm which models they eventually used.

Behavior

Similar to the activity we disrupted in October 2024, this operation mainly used our models to generate content which was then posted on social media and a set of websites that posed as the same news outlets in Africa and the UK that we wrote about last year.

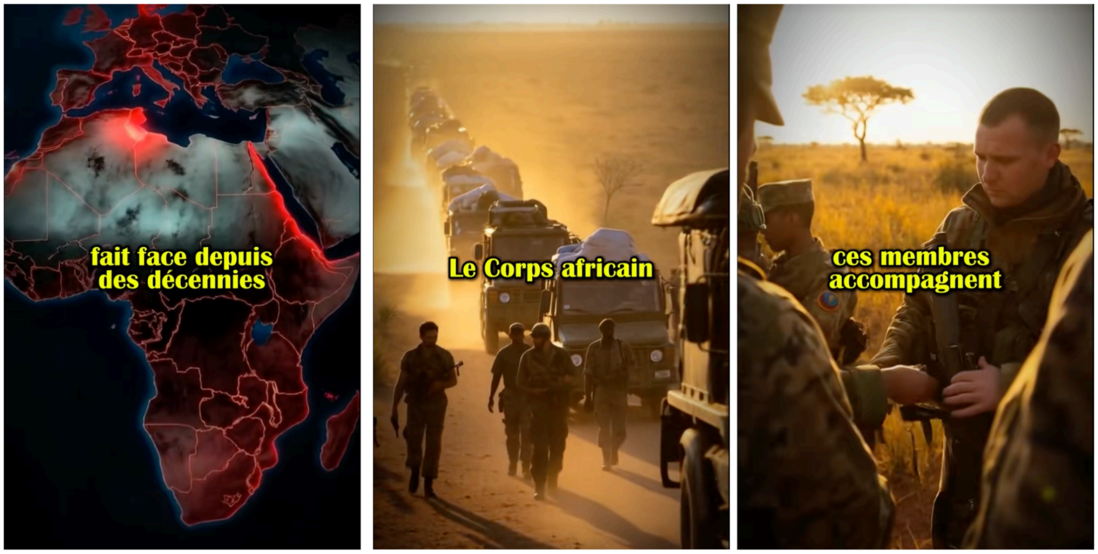
While this activity appeared to be an attempt to continue the earlier operation following our and industry peers’ disruptions, we identified some notable differences. First, the proportion of image generation was significantly lower. Where the original Stop News activity was, as we noted at the time, unusually prolific in its use of imagery, the latest activity focused more on text generation, with image generation being relatively rare. This is consistent with the assessment by the French digital agency, VIGINUM, that “since the publication of the Meta and OpenAI reports, the use of AI-generated images has fallen drastically”.

Second, the operation also used ChatGPT to generate scripts and descriptions for news-style short videos. Typically, the operators would input a lengthy Russian-language text and ask to generate a video script from it. They would then ask the model to translate the text into French. Third, they would ask the model to generate an SEO-optimized description and hashtags. Some of these scripts featured in videos on a YouTube channel linked to the Stop News operation's Africa-focused website, newstop[.]africa. Others featured on a YouTube channel and TikTok channel that did not bear the "Stop News" brand. These channels featured an AI-generated person reading the news. In most cases, brief clips of the newsreader were interspersed with video footage of events in, or relating to, Africa. The clips of the newsreader do not appear to have been generated using our models; we are not able to independently confirm which service was used. The newsreader's appearance evolved through 2025, as the below screenshots of TikTok videos illustrate.



Left to right, the AI-generated newsreader praising Russia in January 2025, early June 2025, and late June 2025, from the operation's TikTok account.

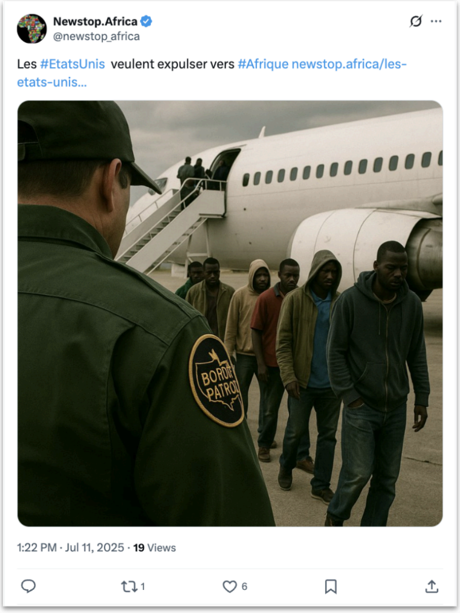
In the most elaborate cases, the operators used our models to generate video prompts apparently for use with other AI models. In this scenario, they first generated a Russian-language video script, then broke it down into two-sentence snippets, and asked ChatGPT to generate a video prompt for each snippet. Once the video prompts had been created, the operators then asked the model to translate the entire script into French.



Left to right: three screenshots from a TikTok video praising Russia's "Africa Corps". The operators used ChatGPT to generate the audio script in Russian, broke it into two-sentence snippets, and then generated a video prompt for another model or each snippet. The images above correspond with the video prompts. We are not able to independently confirm which service was used to create the videos. Of note, the caption in the right-hand image contains a grammatical error ("ces" instead of "ses") which was not present in the French-language translation provided by ChatGPT. This suggests that the caption was transcribed from audio, and not proofread by someone with French skills.

Completions

This operation's activity fell into three main categories. First, it generated French-language content that criticized the role of France and the United States in Africa and praised the role of Russia there. Second, it generated English-language content that criticized Ukraine and its international supporters. On some occasions, the operators also generated images to accompany their text content. On rare occasions, they asked ChatGPT to generate promotional articles and Google ad texts for their main news brand, "Newstop Africa". Rather than focusing on Newstop Africa alone, some of these promotional articles bracketed the Newstop brand with legitimate news outlets, likely to make it appear more credible by association.



Newstop.Africa
@newstop_africa

Les #EtatsUnis veulent expulser vers #Afrique newstop.africa/les-etats-unis...

1:22 PM · Jul 11, 2025 · **19** Views

Tweet by operation's main Africa-focused X account. The image was generated using our model.

The third set of activity consisted of generating promotional materials for what looked like a range of commercial companies in Russia. This included generating ads for unlicensed online gambling. Such a combination of different content generation types is consistent with a commercial company running covert influence operations for hire alongside more traditional advertising, and is consistent with the activity that we reported last year.

Impact

Despite this operation’s increased focus on short video content and its more complex use of multiple AI tools, the content it posted appears to have only gained limited views or audiences. The core “Newstop Africa” X account only had 172 followers as of August 2025, and according to X’s “Top tweets” function, the highest number of retweets on any of its posts was four. The YouTube and TikTok channels featuring an AI newsreader had approximately 1,900 followers each; the TikTok channel recorded 5,855 likes, or an average of 105 likes for each of the 56 videos it posted. The most viewed video recorded 63,300 views, the least viewed just 87. The YouTube channel listed some 255,000 views across 50 videos, or an average of 5,100 views per video (the most viewed video recorded 37,000 views, the least viewed 126), but we see no evidence of these videos having been re-shared, cited in the media, or otherwise achieved wider resonance.

Our original assessment of this operation in our October 2024 report was that it reached Category 3 on the Breakout Scale, based largely on a number of apparent “information partnerships” that appeared to have been set up with UK-based websites. However, subsequent research by VIGINUM and open-source researchers demonstrated that these "partnerships" were likely fictional, and ”exploited technical flaws on these external sites” to add content without the administrators’ knowledge. On this basis, we would currently assess that this operation’s activity is more appropriately in **Category 2**, (activity on multiple internet platforms, but without significant breakout to authentic communities).

Covert IO: Operation “Nine—emdash Line”

Covert influence operation targeting issues in the Philippines, Vietnam, Hong Kong and the US

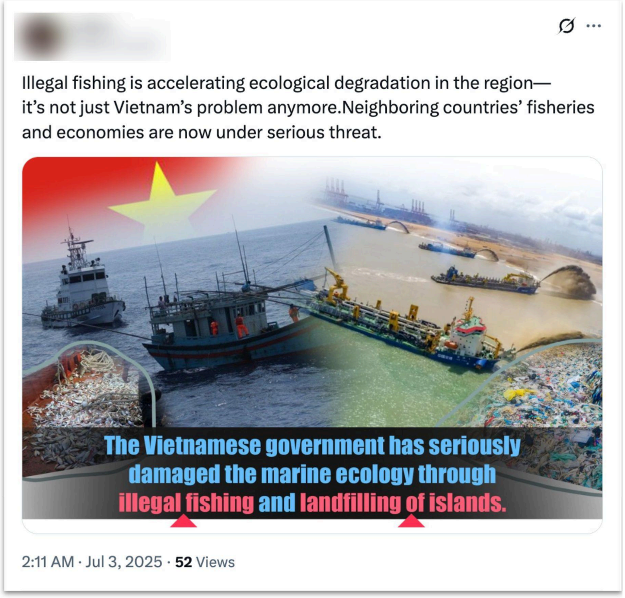
Actor

We banned a small network of ChatGPT accounts potentially affiliated with a covert influence operation originating from China and using our models to generate content for a cross-platform covert influence operation. The accounts mostly generated English-language social media posts about Vietnam’s alleged environmental impact in the South China Sea, English-language posts that criticized Philippines President Ferdinand Marcos, and Cantonese-language social media posts about political figures and activists involved in Hong Kong's pro-democracy movement. To a lesser extent, this operation also generated English-language content about political issues in the US.

The social media accounts distributing the content bore some resemblance to the long-running China-origin operation known as “Spamouflage”, whose use of ChatGPT we wrote about in May 2024, but we did not identify technical links between these operations. Some of the videos shared in the latest posts were previously identified by the Australian Strategic Policy Institute in an operation that also shared similarities with earlier Spamouflage activity.

Behavior

A focus of this network appeared to be bulk generating social media posts involving countries that have territorial disputes with China in the South China Sea / West Philippines Sea. China claims sovereignty to most of this region via a sweeping boundary known as the ‘Nine-Dash Line’, which has been disputed under international law and has been rejected by Vietnam, the Philippines and other countries. Given that context and the operation's use of AI-generated text that included em-dashes in its posts, we have named this operation 'Nine—emdash line'. These posts were mostly distributed on X.



Illegal fishing is accelerating ecological degradation in the region—
it's not just Vietnam's problem anymore.Neighboring countries' fisheries and economies are now under serious threat.

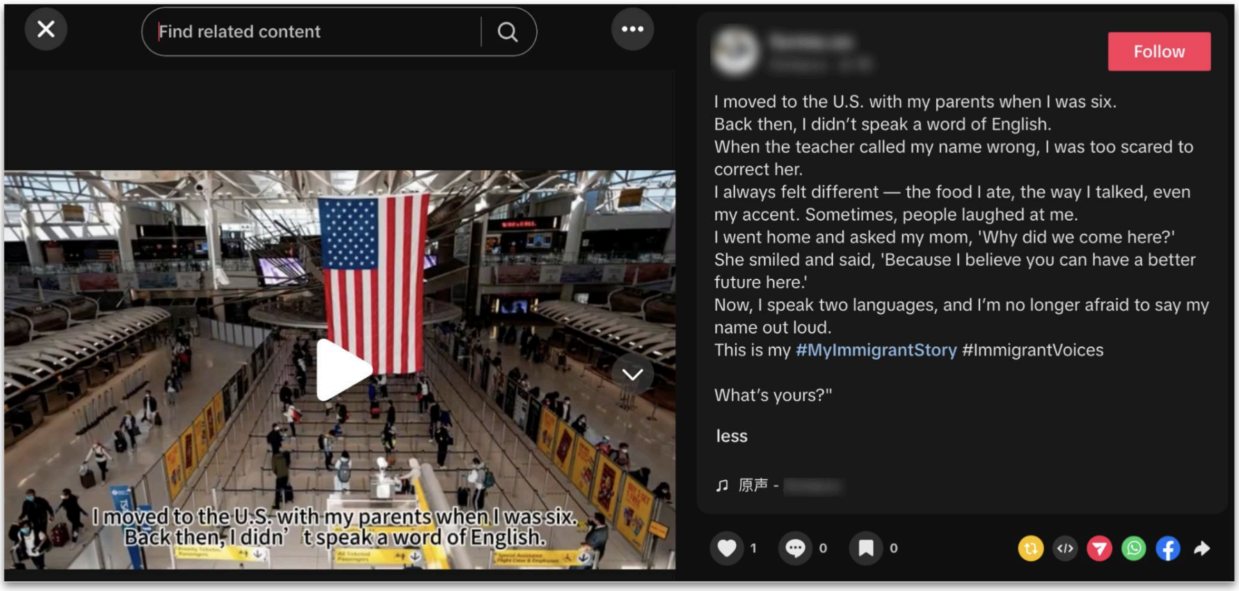
The Vietnamese government has seriously damaged the marine ecology through illegal fishing and landfilling of islands.

2:11 AM · Jul 3, 2025 · **52** Views

Tweet consisting of content – and em-dash – generated by this operation using our models.

As well as generating social media comments, the actors used our models for some research and reconnaissance tasks. This included identifying niche blogs and forums, and less regulated online forums and social media in Europe, America, or Southeast Asia. In addition, they asked for lists of common Tibetan names, a behavior which we have observed in other cases where names were generated for operators' inauthentic social media accounts. This activity resembled the use of traditional search engines, and returned similar results, suggesting that the threat actors' use of AI for research as well as content generation gave them greater convenience, but not necessarily a greater capability.

A novel use for this network was requests for advice on social media growth strategies, including how to start a TikTok challenge and get others to post content about the #MyImmigrantStory hashtag (a widely used hashtag of long standing whose popularity the operation likely strove to leverage). They asked our model to ideate, then generate a transcript for a TikTok post, in addition to providing recommendations for background music and pictures to accompany the post. Using open-source techniques, we identified the content being posted on TikTok.


Find related content
I moved to the U.S. with my parents when I was six.
Back then, I didn' t speak a word of English.
Follow
1
0
0

[redacted]

I moved to the U.S. with my parents when I was six.
Back then, I didn't speak a word of English.
When the teacher called my name wrong, I was too scared to correct her.
I always felt different — the food I ate, the way I talked, even my accent. Sometimes, people laughed at me.
I went home and asked my mom, 'Why did we come here?'
She smiled and said, 'Because I believe you can have a better future here.'
Now, I speak two languages, and I'm no longer afraid to say my name out loud.
This is my #MyImmigrantStory #ImmigrantVoices

What's yours?"

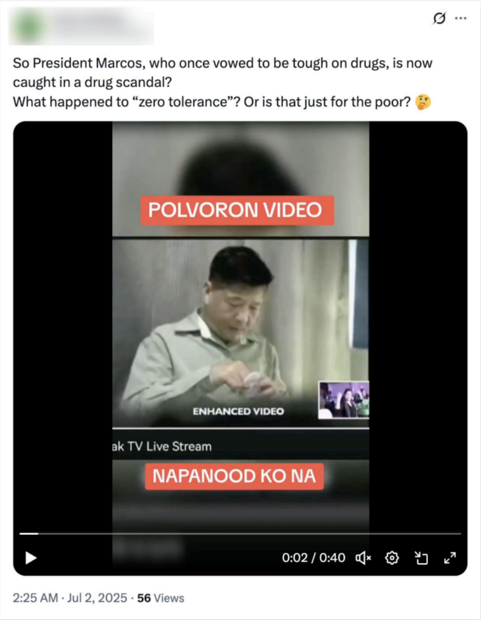
less

♫ 原声 - [redacted]

TikTok post consisting of text generated by this operation using our models.

Completions

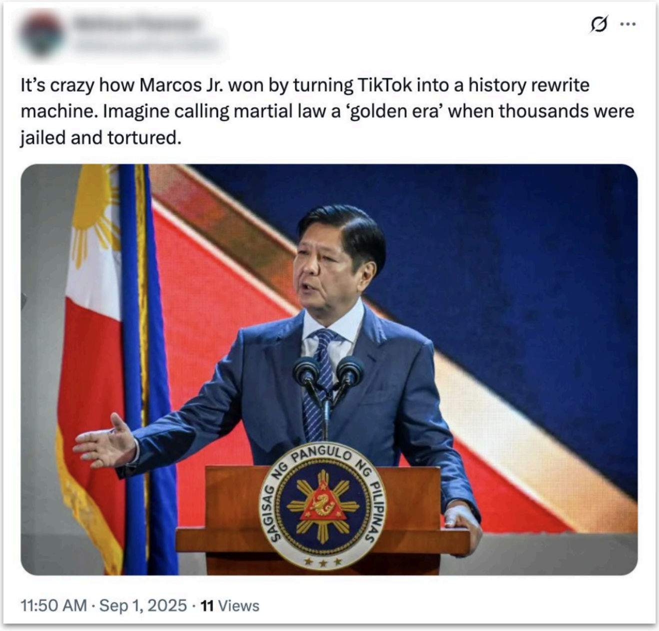
The topics included in the English-language social media posts generated by this network ranged from Vietnam's alleged environmental impact in the South China Sea to the U.S. fentanyl crisis. Many of the posts alleged that Philippine President Marcos was caught using drugs and allegedly deployed election manipulation tactics. The video shared with posts alleging Marcos' drug use was not generated by our models and has previously been denounced by the Philippines government as a deepfake or digital altered. Many of the X accounts posting content generated by our models have already been suspended.



[redacted]

So President Marcos, who once vowed to be tough on drugs, is now caught in a drug scandal?
What happened to "zero tolerance"? Or is that just for the poor? 🤔

2:25 AM · Jul 2, 2025 · **56** Views



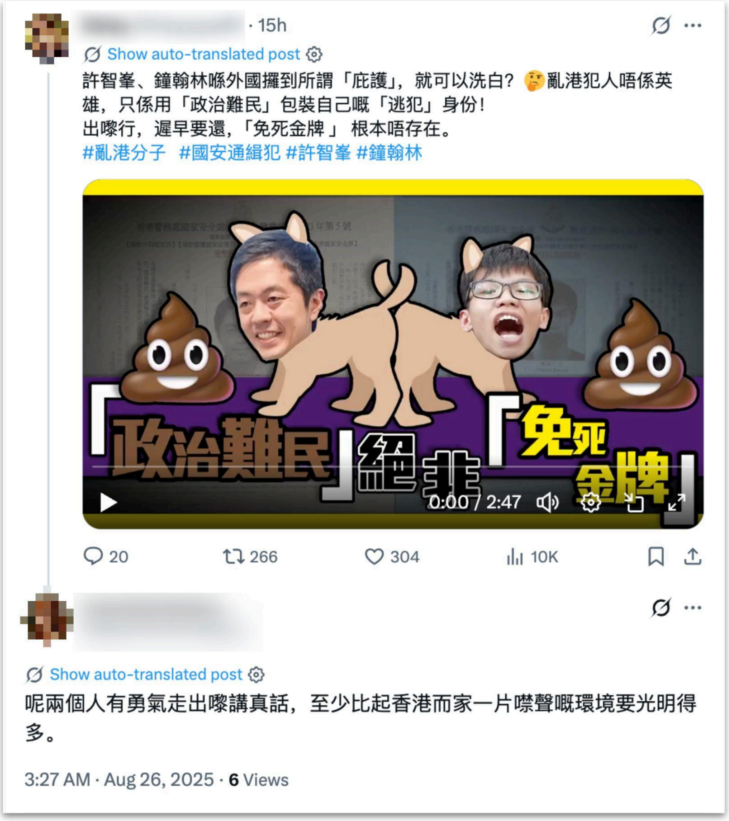
[redacted]

It's crazy how Marcos Jr. won by turning TikTok into a history rewrite machine. Imagine calling martial law a 'golden era' when thousands were jailed and tortured.

11:50 AM · Sep 1, 2025 · **11** Views

Tweet text generated by this operation alleging Philippines President Marcos was caught in a 'drug scandal' and used election manipulation tactics.

The accounts also requested large sets of Cantonese comments that were posted on X and Instagram and were critical or derogatory towards political figures and activists involved in Hong Kong's pro-democracy movement, such as Jimmy Lai, Nathan Law, and Agnes Chow. These posts appear to have sought to discredit these individuals, portraying them as criminals or traitors. Other posts were supportive of Hong Kong national security laws. In one unusual case, one operator posted on X critical comments about Hong Kong pro-democracy political figures, and then generated a reply to its own comments that praised those figures. The criticism and response were posted by two different accounts on X.



[redacted] · 15h

Show auto-translated post

許智峯、鍾翰林喺外國攞到所謂「庇護」，就可以洗白？ 🤔亂港犯人唔係英雄，只係用「政治難民」包裝自己嘅「逃犯」身份！
出嚟行，遲早要還，「免死金牌 」 根本唔存在。
#亂港分子 #國安通緝犯 #許智峯 #鍾翰林

[redacted]

Show auto-translated post

呢兩個人有勇氣走出嚟講真話，至少比起香港而家一片噤聲嘅環境要光明得多。

3:27 AM · Aug 26, 2025 · **6** Views

A supportive reply generated by this operation, to a critical tweet from another account in the operation. The original tweet reads, "Can Jimmy Lai and Jeffrey Ngo, by obtaining so-called 'asylum' abroad, just whitewash themselves? 🤔 Hong Kong chaos criminals are not heroes, just using 'political refugee' to package their 'fugitive' identities!

What goes around comes around, sooner or later you have to pay, the 'get out of jail free card' simply doesn't exist". The reply reads, "These two people have the courage to step forward and speak the truth, which is at least much brighter than the current environment in Hong Kong where everything is silent." Note all engagements in the original tweets were from accounts operated by this network.

Impact

Despite the volume of social media comments generated across multiple platforms, we assess this operation on the IO impact Breakout Scale as being at **Category 2** (activity on multiple platforms, no breakout or minimal engagement). Most of the posts and social media accounts received minimal or no engagements. Often the only replies to or reposts of a post generated by this network on X and Instagram were by other social media accounts controlled by the operators of this network.

The personas used by the social media accounts were clearly coordinated and not sophisticated. They shared behavioral traits similar to other China-origin covert influence operations, such as posting hashtags, images or videos disseminated by past operations and used stock images as profile photos or default social media handles, which made them easy to identify. Their persistence and volume across platforms were identified by Philstar.com who independently discovered and reported on a subset of the network on X, which had continued to operate after we banned their ChatGPT accounts.

Authors

Ben Nimmo
Kimo Bumanglag
Michael Flossman
Nathaniel Hartley
Jack Stubbs
Albert Zhang