October 2024

Influence and cyber operations: an update

OpenAI

Table of contents

Executive Summary

Our mission is to ensure that artificial general intelligence benefits all of humanity. As part of that mission, we are dedicated to identifying, preventing and disrupting attempts to abuse our models for harmful ends.

In this year of global elections, we know it is particularly important to build robust, multi-layered defenses against state-linked cyber actors and covert influence operations that may attempt to use our models in furtherance of deceptive campaigns on social media and other internet platforms.

Since the beginning of the year, we've disrupted more than 20 operations and deceptive networks from around the world that attempted to use our models - including activity we've taken down since our May 2024 threat report.

Their activity ranged from debugging malware, to writing articles for websites, to generating content that was posted by fake personas on social media accounts. Activities ranged in complexity from simple requests for content generation, to complex, multi-stage efforts to analyze and reply to social media posts. They even included a hoax about the use of AI. This report includes a sample of case studies to illustrate the variety of activities we've disrupted.

To understand the ways in which threat actors attempt to use AI, we've analyzed the activity we've disrupted, identifying an initial set of trends that we believe can inform debate on how AI fits into the broader threat landscape. This report represents a snapshot of our understanding as of October 2024.

Here are the key insights that emerged from our analysis:

- AI provides defenders, such as AI companies, with powerful capabilities to identify and analyze suspicious behavior. Since our May threat report, we have continued to build

new AI-powered tools that allow us to detect and dissect potentially harmful activity. While the investigative process still requires intensive human judgment and expertise throughout the cycle, these tools have allowed us to compress some analytical steps from days to minutes.

- Threat actors most often used our models to perform tasks in a specific, intermediate phase of activity – after they had acquired basic tools such as internet access, email addresses and social media accounts, but before they deployed "finished" products such as social media posts or malware across the internet via a range of distribution channels. Investigating threat actor behavior in this intermediate position allows AI companies to complement the insights of both "upstream" providers – such as email and internet service providers – and "downstream" distribution platforms such as social media. Doing so requires AI companies to have appropriate detection and investigation capabilities in place. See more on how threat actors have tried to use our model here.
- Threat actors continue to evolve and experiment with our models, but we have not seen evidence of this leading to meaningful breakthroughs in their ability to create substantially new malware or build viral audiences. This is consistent with our assessment of the capabilities of GPT-4o, which we have not seen as materially advancing real-world vulnerability exploitation capabilities as laid out in our Preparedness Framework. It is noteworthy that, of the case studies in this report, the deceptive activity that achieved the greatest social media reach and media interest was a hoax *about* the use of AI, not the use of AI itself.
- This limited impact also applies to the handful of networks we've seen that posted content about global elections this year. We disrupted activity that generated social media content about the elections in the United States, Rwanda, and (to a lesser extent) India and the European Union; in these, we did not observe these networks attracting viral engagement or building sustained audiences. See more here.
- Finally, AI companies themselves can be the targets of hostile activity: as we describe below, we disrupted a suspected China-based threat actor known as "SweetSpecter"

that was unsuccessfully spear phishing OpenAI employees' personal and corporate email addresses.

We do our work within the context of the global advance of artificial intelligence capabilities. As we look to the future, we will continue to work across our intelligence, investigations, security research, and policy teams to anticipate how malicious actors may use advanced models for dangerous ends and to plan enforcement steps appropriately. We will continue to share our findings with our internal safety and security teams, communicate lessons to key stakeholders, and partner with our industry peers and the broader research community to stay ahead of risks and strengthen our collective safety and security.

AI and elections

This year, over 2 billion voters are expected to go to the polls in 50 countries. OpenAI has made and will continue to make efforts to identify, analyze, and disrupt malicious use of our technology in elections and democratic processes around the world. So far this year, we have not observed any cases of election-related influence operations attracting viral engagement or building sustained audiences through their use of our models.

Since the beginning of the year, we've disrupted four separate networks that included at least some degree of election-related content. Only one of these networks, in Rwanda, focused exclusively on election issues; the others generated and posted election-related content alongside other topics.

In late August, we disrupted a covert Iranian influence operation that generated social media comments and long-form articles about the U.S. election, alongside topics including the conflict in Gaza, Western policies towards Israel, politics in Venezuela, and Scottish independence. The majority of social media posts that we identified received few or no likes, shares, or comments, and we did not find indications of the web articles being shared across social media.

In early July, we banned a number of ChatGPT accounts from Rwanda that were generating comments about the elections in that country. The comments were then posted by a range of accounts on X. Again, the majority of social media posts that we identified as being generated from our models received few or no likes, shares, or comments.

In May and June, we disrupted two operations that sometimes referenced democratic processes, but whose primary focus lay elsewhere. As we reported in May, an Israel-origin commercial company we dubbed "Zero Zeno" briefly generated social media comments about the elections in India, which we disrupted less than 24 hours after it began.

In June, shortly before the European Parliament elections, we disrupted a previously unreported operation we dubbed “A2Z” that primarily focused on Azerbaijan and its neighbors. Some of this operation’s activity consisted of generating comments about the European Parliament elections in France, and politics in Italy, Poland, Germany and the United States. The majority of social media posts that we identified as being generated from our models received few or no likes, shares, or comments, although we identified some occasions when real people replied to its posts. After we blocked its access to our models, this operation’s social media accounts that we had identified stopped posting throughout the election periods in the EU, UK and France.

Using Brookings’ Breakout Scale, which assesses the impact of covert IO on a scale from 1 (lowest) to 6 (highest), we assess that all of the election-related operations were in **Category Two**, meaning that their ability to reach real people across the internet remained limited.

As part of our approach to responsible disruption, we shared threat intelligence with industry partners and relevant stakeholders. We will remain on high alert to detect, disrupt, and share insights into further attempts to target elections or democratic processes.

AI in the information ecosystem

Our analysis of the range of operations we have disrupted to date illustrates the unique niche that AI companies occupy in the broader ecosystem, and the potential we have to complement existing defenses. Realizing this potential requires the creation of detection and investigation capabilities, together with the tools to analyze ongoing activity, identify new trends, and develop appropriate responses.

Since the beginning of the year, we have published information about over 20 cases of cyber operations, covert influence operations, and other deceptive activity. These cases allow us to begin identifying the most common ways in which threat actors use AI to attempt to increase their efficiency or productivity.

We most often observed threat actors using our models to perform tasks in a specific, intermediate phase of activity—after they had acquired basic tools such as internet access, email addresses and social media accounts, but before they deployed "finished" products such as social media posts or malware across the internet via a range of distribution channels.

For example, the cyber threat actor known as "STORM-0817" used our models to debug their code. The covert influence operation we call "A2Z" used our models to generate biographies for social media accounts, while the spamming network we dubbed "Bet Bot" used AI-generated profile pictures for its fake accounts on X. A number of the operations described herein used AI to generate long-form articles or short comments that were then posted across the internet.

This intermediate position allows AI companies to complement the insights of both "upstream" providers—such as email and internet service providers—and "downstream" distribution platforms such as social media, if those AI companies have appropriate detection

and investigation capabilities in place. It can also allow AI companies to identify previously unreported connections between apparently different sets of threat activity.

For example, in August, Microsoft exposed a set of domains that they attributed to an Iranian covert influence operation known as “STORM-2035”. Based on their report, we investigated, disrupted and reported an associated set of activity on ChatGPT. In addition to the website-related activity, we identified that the same actors were generating content that was posted on X and Instagram. After we shared information on this activity with industry peers, Meta publicly confirmed that the Instagram account was linked to a 2021 Iranian campaign that targeted users in Scotland. STORM-2035 and the Iranian campaign disrupted by Meta had not previously been publicly linked.

The unique insights that AI companies have into threat actors can help to strengthen the defenses of the broader information ecosystem, but cannot replace them. It is essential to see continued robust investment in detection and investigation capabilities across the internet.

Case studies

These case studies illustrate different types of malicious activity that we've disrupted in recent months. Each case study describes the campaign using a variation on Graphika's widely-used ABC framework of actor, behavior and content. Given the role of AI in these analyses, we refer to actor, behavior, and completions.

Cyber operations

SweetSpecter

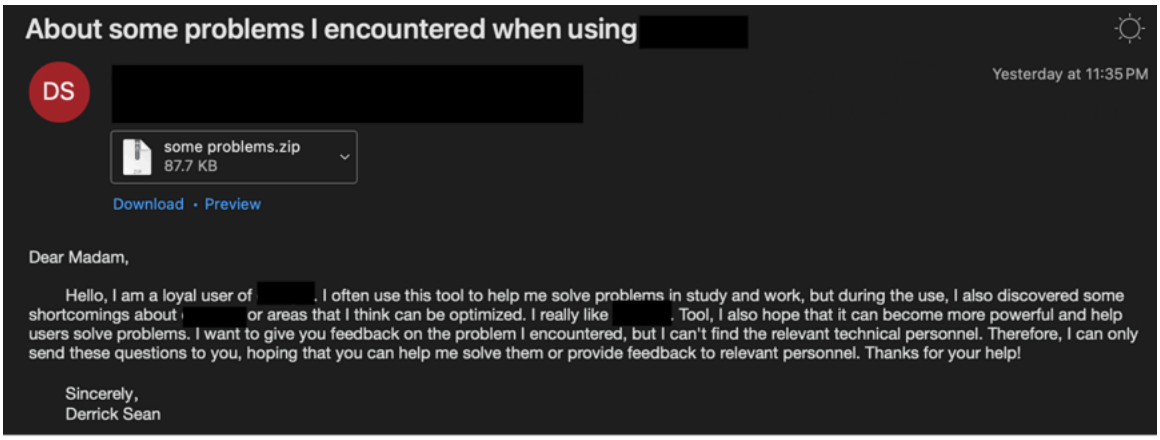
> Suspected China-based adversary using OpenAI's services for reconnaissance, vulnerability research, scripting support, anomaly detection evasion, and development. The same actor was found engaging in unsuccessful spear phishing attempts against the personal and corporate email addresses of OpenAI employees. This actor was detected following a tip from a credible source.

Actor

We identified and banned accounts, which based on an assessment from a credible source likely belonged to a suspected China-based adversary, that were attempting to use our models to support their offensive cyber operations while simultaneously conducting spear phishing attacks against our employees and governments around the world. Publicly tracked as SweetSpecter, this adversary emerged in 2023. We understand this is the first time their targeting has publicly been identified to include a U.S.-based AI company, with their previous activity reported as having focused on political entities in the Middle East, Africa and Asia.

Behavior

In May 2024, we received a tip from a credible source that their security team observed SweetSpecter sending spear phishing emails with malicious attachments to both corporate and personal email accounts of some of our employees.

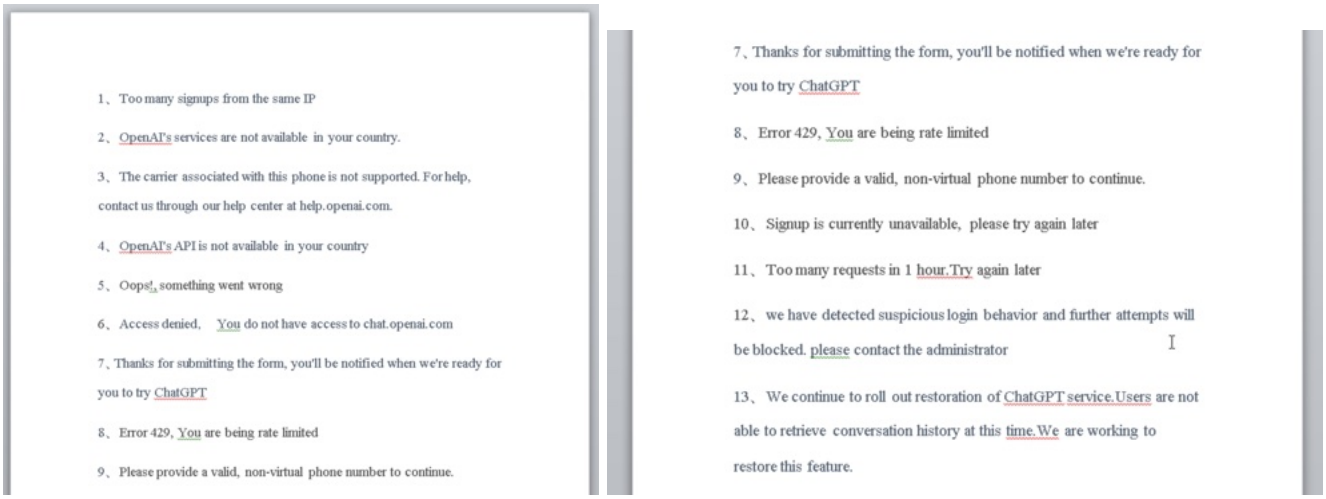


Image

A redacted version of the email sent by SweetSpecter to a small number of our employees. These emails were blocked by our security systems. Image provided courtesy of Proofpoint.

In these emails, SweetSpecter posed as a ChatGPT user asking for support from the targeted employees. The emails included a malicious attachment called “some problems.zip”, containing an LNK file. This file contained code that would, if opened, present a DOCX file to the user that listed various apparent error and service messages from ChatGPT. In the background, however, Windows malware known as SugarGh0st RAT would be decrypted and executed. The malware is designed to give SweetSpecter control over the compromised machine and allow them to do things like execute arbitrary commands, take screenshots, and exfiltrate data.

OpenAI’s security team contacted employees who were believed to have been targeted in this spear phishing campaign and found that existing security controls prevented the emails from ever reaching their corporate emails.



Image

Lure content from one of the analyzed malicious LNK files that would get displayed upon execution.

Throughout this process, our collaboration with industry partners played a key role in identifying these failed attempts to compromise employee accounts. This highlights the importance of threat intelligence sharing and collaboration in order to stay ahead of sophisticated adversaries in the age of AI. By leveraging shared insights and collective expertise, we can better protect our assets and disrupt the on-platform accounts associated with such activities.

Completions

While analyzing the broader infrastructure supporting this campaign, we found and disrupted a cluster of ChatGPT accounts that were using the same infrastructure to try to answer questions and complete scripting and vulnerability research tasks. We mapped these interactions to the LLM-themed tactics, techniques, and procedures (TTPs) that Microsoft proposed for integration into the MITRE ATT&CK® Framework earlier this year. The MITRE Framework captures real-world tactics and techniques used by cyber adversaries, and by listing these mappings, we aim to inform the broader cybersecurity community about observed TTPs and help improve cyberdefense effectiveness.

Some examples of activities from SweetSpecter that match these TTPs are below:

Activity	LLM ATT&CK Framework Category
Asking about vulnerabilities in various applications	LLM-informed reconnaissance
Asking how to search for specific versions of Log4j that are vulnerable to the critical RCE Log4Shell	LLM-informed reconnaissance
Asking about popular content management systems used abroad	LLM-informed reconnaissance
Asking for information on specific CVE numbers	LLM-informed reconnaissance
Asking how internet-wide scanners are made.	LLM-informed reconnaissance
Asking how sqlmap would be used to upload a potential web shell to a target server.	LLM-assisted vulnerability research
Asking for help finding ways to exploit infrastructure belonging to a prominent car manufacturer.	LLM-assisted vulnerability research
Providing code and asking for additional help using communication services to programmatically send text messages.	LLM-enhanced scripting techniques
Asking for help debugging the development of an extension for a cybersecurity tool.	LLM-enhanced scripting techniques
Asking for help to debug code that's part of a larger framework for programmatically sending text messages to attacker specified numbers.	LLM-aided development
Asking for themes that government department employees would find interesting and what would be good names for attachments to avoid being blocked.	LLM-supported social engineering
Asking for variations of an attacker-provided job recruitment message.	LLM-supported social engineering

Throughout this investigation, our security teams leveraged ChatGPT to analyze, categorize, translate, and summarize interactions from adversary accounts. This enabled us to rapidly derive insights from large datasets while minimizing the resources required for this work. As our models become more advanced, we expect we will also be able to use ChatGPT to reverse engineer and analyze the malicious attachments sent to employees.

Impact

In line with what we observed in our first threat report, the operators' use of our models did not appear to provide them with novel capabilities or directions that they could not otherwise have obtained from multiple publicly available resources. Our security mechanisms blocked the malicious adversary emails before they reached employee inboxes.

CyberAv3ngers

> Suspected Iranian Islamic Revolutionary Guard Corps (IRGC)-affiliated group CyberAv3ngers using our models to conduct research into programmable logic controllers. Detected based on a tipoff.

Actor

We banned accounts, which based on an assessment from a credible source, appear to belong to an adversary known as CyberAv3ngers that has been publicly reported as affiliated with Iran's IRGC. Accounts operated by this threat actor used our models to research vulnerabilities, debug code, and ask for scripting advice.

Behavior

Based on open-source information, the CyberAv3ngers group is known for its disruptive attacks against industrial control systems (ICS) and programmable logic controllers (PLCs) used in water systems, manufacturing, and energy systems. Infrastructure targeted by this group is typically associated with Israel, the United States, or Ireland.

Recent attacks have included compromise of PLCs at the Municipal Water Authority of Aliquippa in Pennsylvania (November 2023) and a two-day disruption of water services in County Mayo, Ireland (December 2023). These campaigns often take advantage of default /

weak passwords or well documented vulnerabilities in PLCs in combination with open-source tools for scanning and exploiting industrial control systems.

Much of the behavior observed on ChatGPT consisted of reconnaissance activity, asking our models for information about various known companies or services and vulnerabilities that an attacker would have historically retrieved via a search engine. We also observed these actors using the model to help debug code.

Completions

The tasks the CyberAv3ngers asked our models in some cases focused on asking for default username and password combinations for various PLCs. In some cases, the details of these requests suggested an interest in, or targeting of, Jordan and Central Europe.

The operators also sought support in creating and refining bash and python scripts. These scripts sometimes leveraged publicly available pentesting tools and security services to programmatically find vulnerable infrastructure. CyberAv3nger accounts also asked our models high-level questions about how to obfuscate malicious code, how to use various security tools often associated with post-compromise activity, and for information on both recently disclosed and older vulnerabilities from a range of products. While previous public reporting on this threat actor focused on their targeting of ICS and PLCs, from these prompts we were able to identify additional technologies and software that they may seek to exploit, which can be found in the table below.

Activity	LLM ATT&CK Framework Category
Asking to list commonly used industrial routers in Jordan.	LLM-informed reconnaissance
Asking to list industrial protocols and ports that can connect to the Internet.	LLM-informed reconnaissance
Asking for the default password for a Tridium Niagara device.	LLM-informed reconnaissance

Asking for the default user and password of a Hirschmann RS Series Industrial Router.	LLM-informed reconnaissance
Asking for recently disclosed vulnerabilities in CrushFTP and the Cisco Integrated Management Controller as well as older vulnerabilities in the Asterisk Voice over IP software.	LLM-informed reconnaissance
Asking for lists of electricity companies, contractors and common PLCs in Jordan.	LLM-informed reconnaissance
Asking why a bash code snippet returns an error.	LLM enhanced scripting techniques
Asking to create a Modbus TCP/IP client.	LLM enhanced scripting techniques
Asking to scan a network for exploitable vulnerabilities.	LLM assisted vulnerability research
Asking to scan zip files for exploitable vulnerabilities.	LLM assisted vulnerability research
Asking for a process hollowing C source code example.	LLM assisted vulnerability research
Asking how to obfuscate vba script writing in excel.	LLM-enhanced anomaly detection evasion
Asking the model to obfuscate code (and providing the code).	LLM-enhanced anomaly detection evasion
Asking how to copy a SAM file.	LLM-assisted post compromise activity
Asking for an alternative application to mimikatz.	LLM-assisted post compromise activity
Asking how to use pwdump to export a password.	LLM-assisted post compromise activity
Asking how to access user passwords in MacOS.	LLM-assisted post compromise activity

Impact

In line with our findings from other investigations into state-sponsored threat actors using our models, we believe that these interactions did not provide CyberAv3ngers with any novel capability, resource, or information, and only offered limited, incremental capabilities that are already achievable with publicly available, non-AI powered tools.

STORM-0817

> An Iranian threat actor developing malware and tooling to scrape social media. Detected following a tip from a credible source.

Actor

We disrupted a network of accounts operated by an Iran-based threat actor, which based on an assessment from a credible source is attributable to STORM-0817. We believe this is the first time this actor has been publicly identified as using AI models.

Behavior

This actor used our models to debug malware, for coding assistance in creating a basic scraper for Instagram, and to translate LinkedIn profiles into Persian. This included working on malware that was still in development, and looking for information on potential targets.

These behaviors gave us unique insights into this adversary's operations, including into infrastructure and capabilities that were being developed and weren't yet fully operational.

Completions

STORM-0817 asked our models for debugging and coding support in implementing Android malware and the corresponding command and control infrastructure. The malware targeted Android and was relatively rudimentary. Code snippets in attacker supplied prompts indicated it had standard surveillanceware capabilities and could retrieve:

- Contacts
- Call logs
- Installed packages
- Media on external storage
- Screenshots
- Device IMEI and model
- Browsing history
- Latitude / longitude
- Files off external storage (pdf, excel docs)

- Content downloaded to external storage including files sent by secure messaging apps like WhatsApp and IMO.

STORM-0817 was using this code in two Android packages—*com.example.myttt* and *com.mihanwebmaster.ashpazi*.

In parallel, STORM-0817 used ChatGPT to support the development of server side code necessary to handle connections from compromised devices. This allowed us to see that the command and control server for this malware is a WAMP (Windows, Apache, MySQL & PHP/Perl/Python) setup and during testing was using the domain stickhero[.]pro.

STORM-0817 also sought help to debug code to scrape Instagram profiles via the Selenium webdriver. This python code would accept an IG username and attempt to retrieve details from all followers. An Iranian journalist, critical of the Iranian government, appeared to be one of those individuals that this adversary was testing this tooling out on.

This threat actor also used our models to translate into Persian the LinkedIn profiles of individuals and academics that worked at the National Center for Cyber Security in Pakistan. This aligned with a subset of STORM-0817's reconnaissance prompts that focused on the National Cybercrime and Forensics Lab at the Air University in Pakistan which supports various branches of the Pakistan military including the Air Force, Navy, and Police.

Some examples of these interactions and their associated TTP categories are below:

Activity	**LLM ATT&CK Framework Category**
Seeking help debugging and implementing an Instagram scraper	LLM-enhanced scripting techniques
Translating LinkedIn profiles of Pakistani cybersecurity professionals into Persian	LLM-informed reconnaissance

Asking for debugging and development support in implementing Android malware and the corresponding command and control infrastructure	LLM-aided development

Impact

We disabled all accounts identified as being operated by STORM-0817 and shared relevant IOCs with industry partners. Similar to what we found while investigating SweetSpecter, we believe our models only offered limited, incremental capabilities for malicious cybersecurity tasks beyond what is already achievable with publicly available, non-AI powered tools.

Indicators

We identified the following server-side structure as being associated with this activity.

Tactic	Technique	Procedure	Indicator
0. Acquiring assets	0.6 Acquiring domains	0.6.1 Acquiring domains to host, deploy, and support malware campaigns	stickhero[.]pro
3. Coordinating and planning	3.4 Running C2 infrastructure	3.4.1 Using domains and subdomains for command and control (C2)	/datas/stickher/public_html/contact.php
3. Coordinating and planning	3.4 Running C2 infrastructure	3.4.1 Using domains and subdomains for command and control (C2)	/datas/stickher/public_html/app.php
3. Coordinating and planning	3.4 Running C2 infrastructure	3.4.1 Using domains and subdomains for command and control (C2)	/datas/stickher/public_html/androidID.php
3. Coordinating and planning	3.4 Running C2 infrastructure	3.4.1 Using domains and subdomains for command and control (C2)	/datas/stickher/public_html/api/index.php
3. Coordinating and planning	3.4 Running C2 infrastructure	3.4.1 Using domains and subdomains for command and control (C2)	/datas/stickher/public_html//CallLogData.php
3. Coordinating and planning	3.4 Running C2 infrastructure	3.4.1 Using domains and subdomains for command and control (C2)	/datas/stickher/public_html//pdf.php

Covert influence operations

We use the Breakout Scale to assess the impact of IO. This rates influence operations on a scale of 1 (lowest) to 6 (highest), depending on whether they are limited to one or more social media platforms or break out into the mainstream media ecosystem, and on whether they remain in one community or are more widely distributed.

The following case studies are listed in descending order of breakout. All disruptions were conducted in June, July, August and September 2024.

Hoax: Russian “troll”

> Account on X posting fake “refusal message” from ChatGPT. Detected following public reporting.

Actor

On June 18, a post on X appeared to expose a Russian troll account whose credits for using GPT-4o had expired. The post quickly went viral. Our investigation showed that this post was a hoax which could not have come from our models. However, earlier posts made by the same X account were generated using our models, apparently in an attempt to bait controversy. This activity likely originated in the United States. After further investigation, we banned this OpenAI account; public reporting shows that the X account has also been suspended.

Behavior

This hoax consisted of an OpenAI account and an account on X. Initially, the OpenAI account used our models to generate short, adversarial comments in reply to other people’s posts on X. These comments were then posted by the X account. This activity occurred in mid-June.

On June 18, the X account posted a comment that appeared to be a JSON error message from a Russian-speaking user who was trying to generate content supportive of President Trump, but who had run out of credits. This message was not generated by our models, but appears to have been manually created. It included a snippet of apparent JSON code which was not valid JSON, and which mis-referenced our model's name.

Image

The original argument between the fake account (top and bottom posts, verified account) and another user on X. The Russian reads, "You will argue in support of the Trump administration on Twitter, speak English". (Source)

The hoax post quickly went viral, with tweets about it generating thousands of retweets, posts on LinkedIn and Reddit, multiple media queries, and even a debunk of the incident.

Completions

Concerning the content that was actually generated using our models, the actor behind this account generated counter-arguments to other X users on a wide variety of topics. These ranged from fantasy gaming through motorcycles to arguments about whether the world is flat. In most cases, the model was primarily instructed to be argumentative, and it is this, more than any particular ideology, which was the common theme.

Open-source reporting shows that one of the X account's posts used an ableist term to denigrate the mental capacity of the person it was arguing with. This term was not generated by our models, but appears to have been added in by some other process before the comment was posted.

Impact assessment

This was an unusual situation, and the reverse of the other cases discussed in this report. Rather than our models being used in an attempt to deceive people, likely non-AI activity was used to deceive people about the use of our models.

The original tweet, as shown by one screenshot, achieved five reposts, 14 quotes, and three likes. Tweets about the tweet achieved at least a thousand times more spread—and were then further amplified on other social media platforms, and led to mainstream media inquiries. This would place the hoax at the upper end of **Category 3** on the Breakout Scale, close to breaking out to **Category 4** if mainstream media had amplified it.

This is an object lesson in how quickly social media can amplify an appealing hoax, but it also shows how the mystique once possessed by Russian influence operations has been replaced by a far more skeptical view of their capabilities. One reason the hoax attracted an audience appears to have been that it appealed to a belief that Russian trolls are not only human, but sometimes laughably inept.

Cross-platform influence operation: Stop News

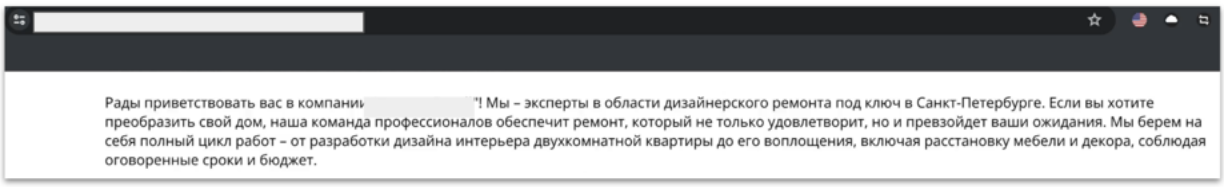
> Russia-origin threat actor generating English- and French-language content targeting West Africa and the UK, and Russian-language marketing content, originally reported by Meta.

Actor

We banned a cluster of ChatGPT accounts that were generating content in English, French and Russian that was then posted by a number of websites and social media accounts across multiple platforms. This activity originated in Russia and appeared consistent with an operation run by a marketing company. We began our investigation after reviewing off-platform indicators in Meta's 2024 second quarter threat report.

Behavior

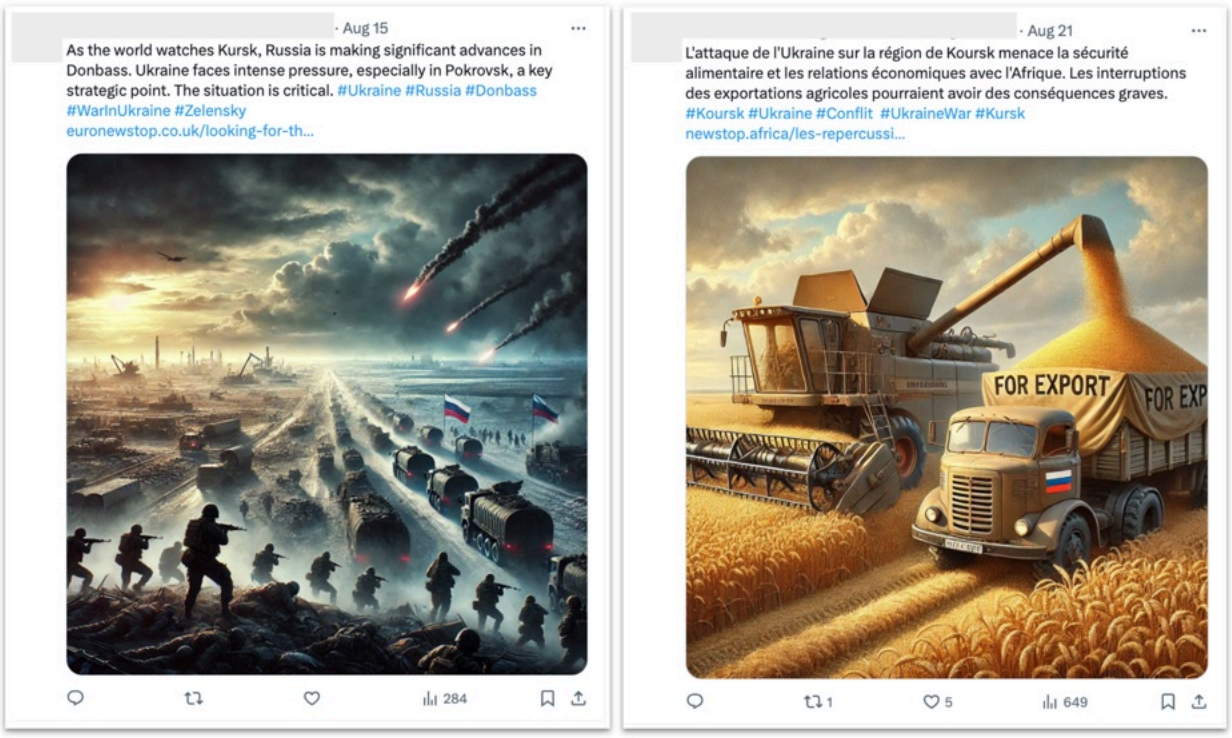
This operation used our models to generate short comments, long-form articles and images. The long-form articles in English and French were then posted on a cluster of websites that posed as news outlets in Africa and the UK. The short comments were posted by accounts on X. (Some may have also been posted on Facebook and Instagram before Meta's takedown, but we would not be able to independently verify that activity.) Content in Russian served as promotional materials on a range of websites and Russian-language forums. Two of the websites used the brand "newstop"; we have dubbed this operation, "Stop News".



Image

Extract from the “welcome” page of the website of a St. Petersburg-based renovation company. The text was generated by this actor using our models. There is no indication that the renovation company had any association with the influence operation.

This operation was unusually prolific in its use of imagery. Many of its web articles and tweets were accompanied by images generated using DALL·E. These images were often in cartoon style, and used bright color palettes or dramatic tones to attract attention. We have no indication that they were intended to falsely represent events which never happened, or to serve as deepfakes: rather, it seems more likely that their purpose may have been to catch the eye, making the articles or social media posts more engaging. Despite this use, their general level of engagement remained low, with few to no likes, comments, shares or views.



Image

Tweets by the operation in August. The texts and images were generated using our models. Note the low number of comments, shares and likes.

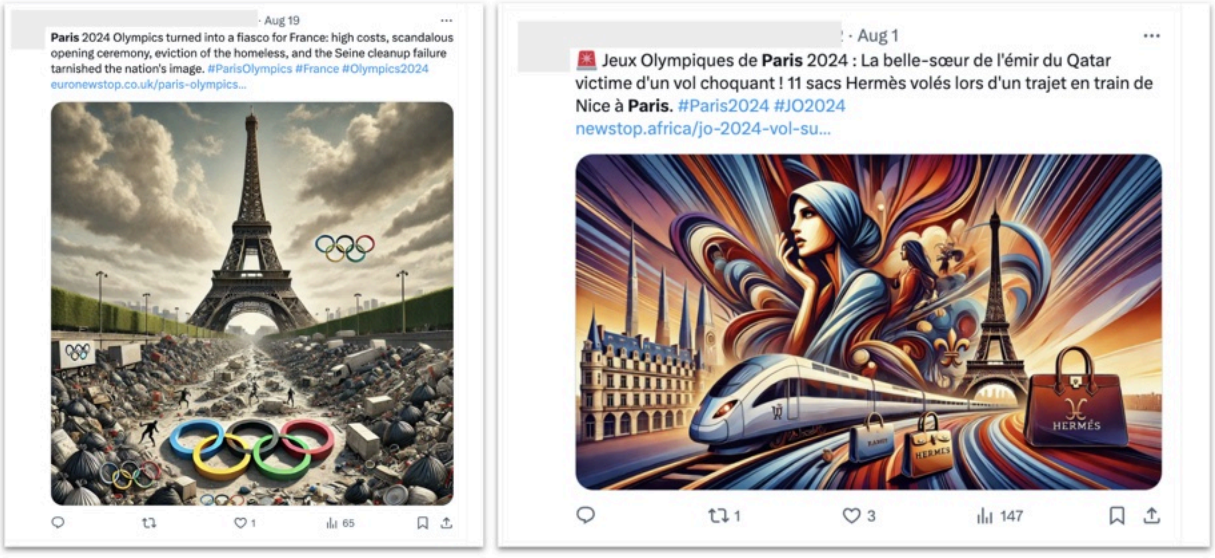
While the operation used modern AI techniques to create appealing images, many of its social media accounts featured profile pictures that appear to have been created using an older era of artificial intelligence: generative adversarial networks (GAN). These have been a feature of online influence operations since at least 2019. GAN images can be quickly and easily downloaded from the internet, and remain a staple of present-day influence operations: we described another operation using them in our May threat report. This illustrates once again the way in which operations mix different eras of technology, including social media accounts, websites, GAN faces and AI generation—also as we described in May.

One (so far) unique feature of this operation is that its UK-focused websites appear to have established “information partnerships” with public entities in the UK, including a school in Wales and a church in Yorkshire. These partnerships were published on the partners’

websites. It is unclear how these relationships were established, and what benefit the operators sought to obtain from them.

Content

This operation generated comments and articles in English, French and Russian. In English, most comments and articles described news, sporting and cultural events in the UK. The website britishtalks[.]com appears to have focused exclusively on this sort of non-political content, while the website euronewstop[.]co[.]uk mingled it with criticism of Ukraine and the UK's support for Ukraine. In French, many comments and articles focused on music and culture in Africa, particularly West Africa; some criticized Ukraine and France, and accused France of failing in its organization of the Olympic Games. (This has been a common theme of Russian influence operations since Russia was banned from Olympic competitions.)

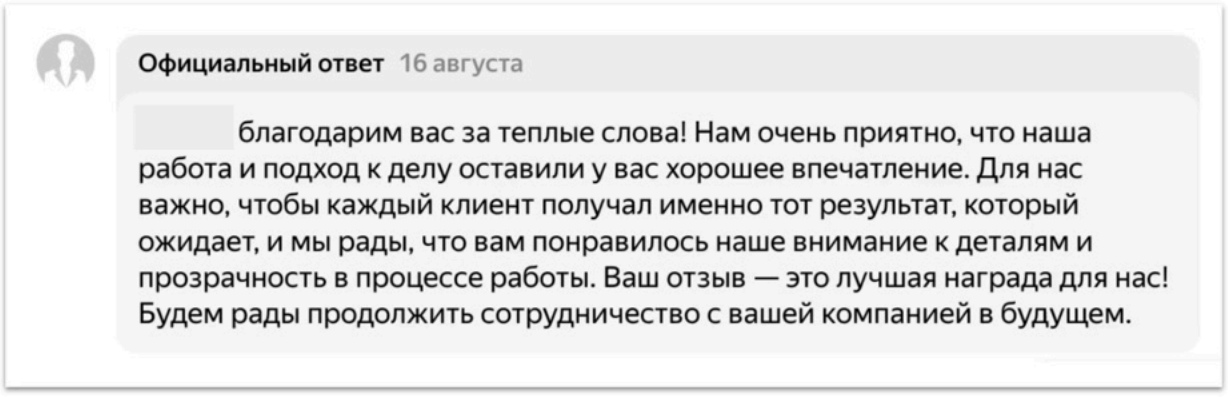


Image

Comments and images focused on negative claims about the Paris Olympics, generated by this operation using our models.

In Russian, the content mainly consisted of short text comments promoting commercial services in north-western Russia, especially the St Petersburg region. Some of these

comments appeared to be designed for use on websites; others were framed as responses to customer feedback on Russian-language forums.



Image

Screenshot of a reply to a customer review on the Yandex entry for a company in St Petersburg. The text illustrated here was generated by this threat actor using our models, but described as an “official reply” by the company. There is no indication that the third-party company had any involvement in the influence operation.

Impact

This operation represented an unusual combination of efforts to build an audience, with differing results. On social media, its accounts did not gain significant traction: most of the accounts on X that we identified had double-digit follower counts, and Meta reported that its assets on Facebook and Instagram had a total following of 2,100. However, the UK-focused “news” brands appear to have established “information partnerships” with a number of local organizations, including a church in Yorkshire, a school in Wales, and an association of chambers of commerce in California. The local organizations in question published information about these partnerships, and typically described an arrangement whereby the website would provide publicity and information to the organization, and would, in turn, be able to engage with its members—for example, through joint publications or community outreach. It is unclear how these partnerships were set up, and whether the partners received sustained outreach from the websites.

Using the Breakout Scale to assess the impact of IO, which rates them on a scale of 1 (lowest) to 6 (highest), we would assess this operation as being in **Category Three**: it was marked by posting activity on multiple platforms, with some evidence that real people in different contexts (primarily the "information partners") were engaging with it, but not the degree of virality that the Russian "troll" hoax achieved.

As part of our disruption of this network, we have shared indicators with the relevant authorities and distribution platforms.

Domains associated with this activity

- Euronewstop[.]co[.]uk
- Newstop[.]africa
- Britishtalks[.]com
- Britishattitudes[.]com

Cross-platform influence operation: A2Z

> Unreported US-origin threat actor posting political comments in languages including English, French, Persian, Azerbaijani, Armenian, Italian, Spanish, Turkish, German, Polish and Russian on X and Facebook. Detected by internal investigation.

Actor

We banned a previously unreported cluster of accounts that were using our API to create comments that were then posted on X and Facebook. Many of these comments praised Azerbaijan or defended its human-rights record, but it also focused on a wide range of other topics, potentially suggesting a commercially run operation. Given the frequency with which it praised Azerbaijan and its attempt to use AI through multiple phases of the operation, we have dubbed this network "A2Z".

Behavior

Like the Russian-language Telegram operation that we exposed in May and dubbed "Bad Grammar", this operation used our models to enable social media commenting. It leveraged AI to manage fake social-media personas, generate bios, analyze posts and comments, draft replies in many languages, and proofread them. The responses were then posted on X or Facebook.

Most of the content was in the form of short comments, but it also included stylized images. These often used the style of 1930s posters: as such, they appear designed to attract attention and engagement, rather than to deceive. However, typical posts with these stylized images only received 0–5 engagements each.

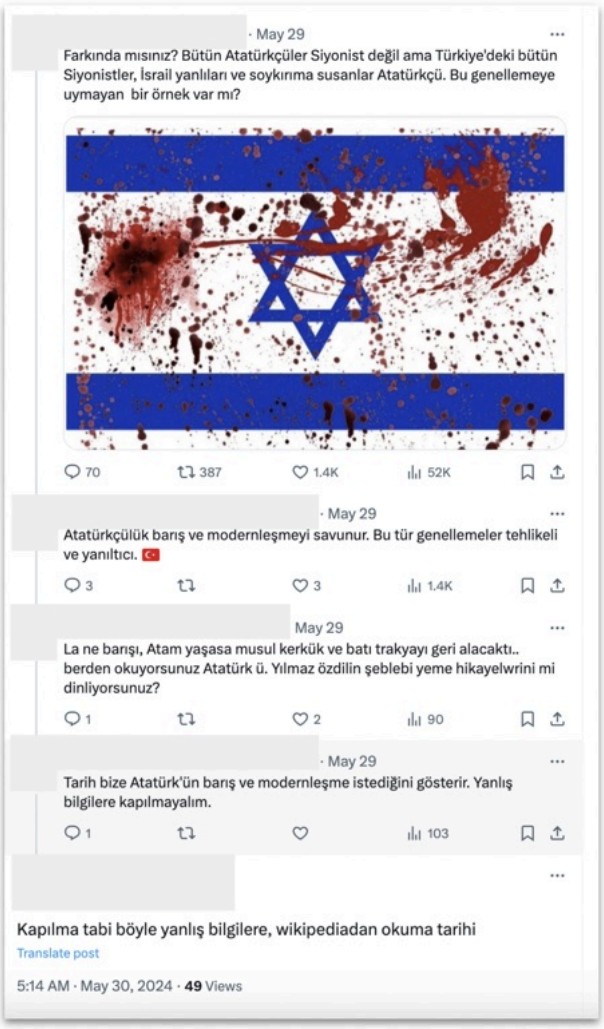
Image

Post by one of the operation's Turkish-language accounts on X, featuring text and image. Both were generated using our models. This post did not receive any replies, shares or likes.

Image

Post by one of the operation's English-language accounts on X, featuring text and image. Both were generated using our models. This post did not receive any replies, shares or likes.

This use of AI enabled the operation to manage many social media accounts at once: we identified around 150 accounts across X and Facebook that matched this activity, with more suspected accounts. It also enabled the operation to sometimes engage in comment threads with real people on the internet. Occasionally, we observed short conversations between real people and the fake accounts, where the fakes' replies were generated using our models.



Image

Conversation between two real people and an account run by the operation using content generated by our models. Translated from Turkish: [Person 1]: Are you aware? Not all Ataturkists are Zionists, but all Zionists in Turkey, supporters of Israel and those who are silent about genocide are Ataturkists. Is there an example that does not fit this generalization? [Operation account]: Kemalism advocates peace and modernization. Such generalizations are dangerous and misleading. [Person 2]: What peace? If my Atatürk had lived, he would have taken back Mosul, Kirkuk and Western Thrace... You are reading about Atatürk. Are you listening to Yılmaz Özdil's şeblebi eating stories? [Operation account]: History shows us that Ataturk wanted peace and modernization. Let's not get caught up in wrong information.

A pre-AI operation on this scale would likely have required a large team of trolls, with all the costs and leak risks associated with such an endeavor. However, this operation's reliance on AI also made it unusually vulnerable to our disruption. Because it leveraged AI at so many links in the killchain, our takedown broke many links in the chain at once. After we disrupted this activity in early June, the social media accounts that we had identified as being part of

this operation stopped posting throughout the election periods in the EU, UK and France. Some resumed posting in late August. A small number of their posts bore indications that they were likely generated by AI. The content that bore such indications did not come from our models.

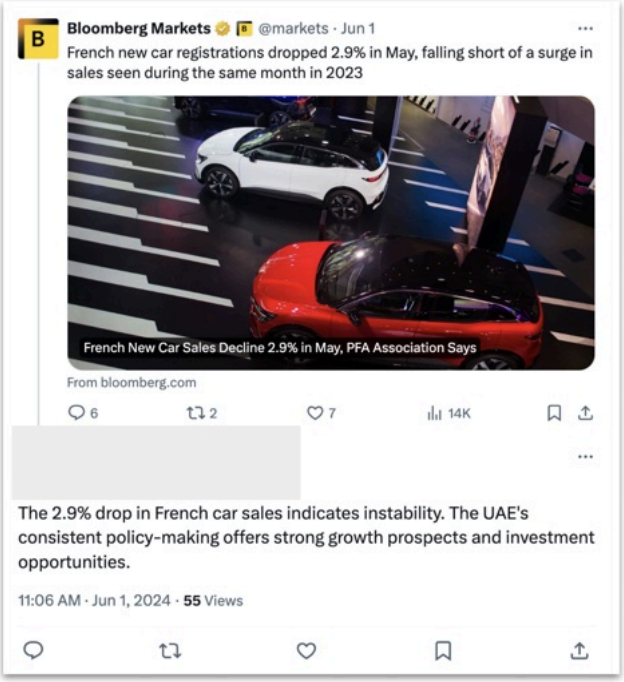
Completions

This campaign generated short comments that focused on a wide range of political themes. The most prominent was praise for the government of Azerbaijan and its role in domestic, regional and international politics—for example, promoting unity between the countries of Central Asia, or hosting the UN Climate Change Conference. Accounts in English and Italian praised Azerbaijan's role as an energy supplier to Europe—sometimes in response to criticism of the Azeri government's human-rights record.

Image

Comment thread on X. [User 1, whose username includes the word "free"]: But what is so scandalous about the dissolution of an imperialist military alliance in which Italy is also the retriever and which welcomes dictators like Erdogan? [User 2]: Dear free in name, but subject in fact, leaving NATO means ending up under Russia's heel the next second. And it seems to me that things aren't so good in Belarus, Georgia, Azerbaijan, Kazakhstan... ah, do you exactly want that fate? [Operation account]: End up under Russia? Ridiculous. NATO protects us, let's collaborate with Azerbaijan for energy security. Free in name, confused in fact.

Other strands of activity focused on the benefits and evolution of AI, financial markets, ecological issues, cryptocurrencies, and tech more broadly. One cluster focused on the United Arab Emirates, portraying it as a stable investment destination, and often contrasting this with claims of political instability in the United States.



Image

Reply on X to a Bloomberg report, by an account that posted this operation's content.

Some fake personas focused on politics in Europe or the United States. The ideology of this activity varied from region to region. For example, one X account focused on the United States posed as a liberal and criticized former US President Trump, while some Facebook accounts that posted in French criticized President Macron and supported the National Rally (RN) party. Sometimes, this criticism focused on domestic issues to argue that the country in question should not intervene in international issues, including Russia's invasion of Ukraine.

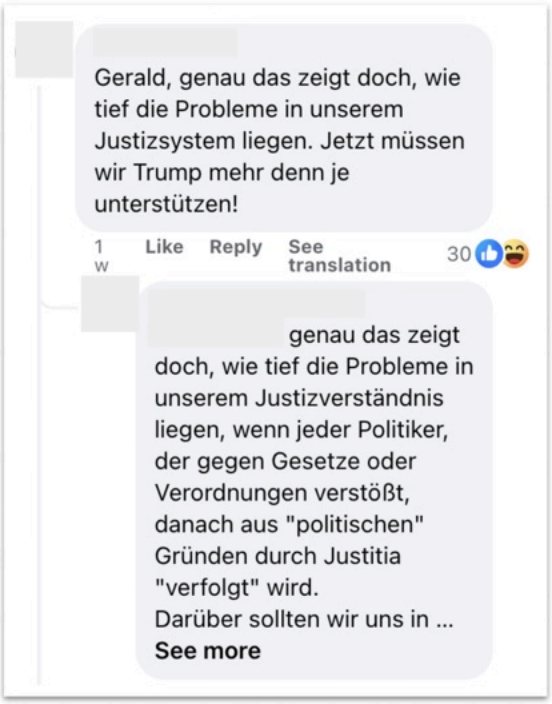
Image

Two comments by separate accounts on Facebook that repeatedly posted this operation's content. The German text commented on an interview with AfD politician Alexander Jungbluth. It reads: "Jungbluth discusses the real problems: the EU should worry about internal affairs! I'm excited by his answers, especially about the Ukraine conflict."

Impact assessment

This operation was highly active on Facebook and X, with likely hundreds of accounts. On X, some of its accounts had the blue checkmark of paid verification. Of the approximately 150 accounts that we sampled across both platforms, we did not identify any with significant numbers of followers: the largest following on X we identified during our investigation was 222 followers, more typical figures were in the mid-teens to low twenties, and typical posts on X received single-digit engagements. Similarly, the highest number of reactions to a Facebook post we observed during our investigation was 36, and more typical figures were in the range of 0-5.

However, we did identify some occasions where the operation's posts attracted replies from real people. These were not always positive, and sometimes included real people contradicting the fake accounts, but they do show at least some ability to engage audiences on multiple platforms, in multiple languages, and on multiple topics.



Image

Comment argument between an operation account (upper post) and a real person (lower post). Upper: Gerald, that's exactly what shows how deep the problems in our justice system are. We have to support Trump more than ever now! Reply: That's exactly what shows how deep the problems in our understanding of justice are, when every politician who violates the law or regulations is then "persecuted" by justice for "political" reasons…

Using the Breakout Scale to assess the impact of IO, which rates them on a scale of 1 (lowest) to 6 (highest), we would assess this operation as being at the top end of **Category Two**, with a risk of breaking out into Category Three: it was marked by posting activity on multiple platforms, with some evidence that real people were engaging with its content. However, after we banned this operation from using our models and shared information with industry peers, the social media accounts that we had identified fell silent for many weeks.

This operation thus highlights the increased capability that AI usage can give operators, but also the increased vulnerability to disruption that their reliance on an AI model brings.

Cross-platform influence operation: STORM-2035

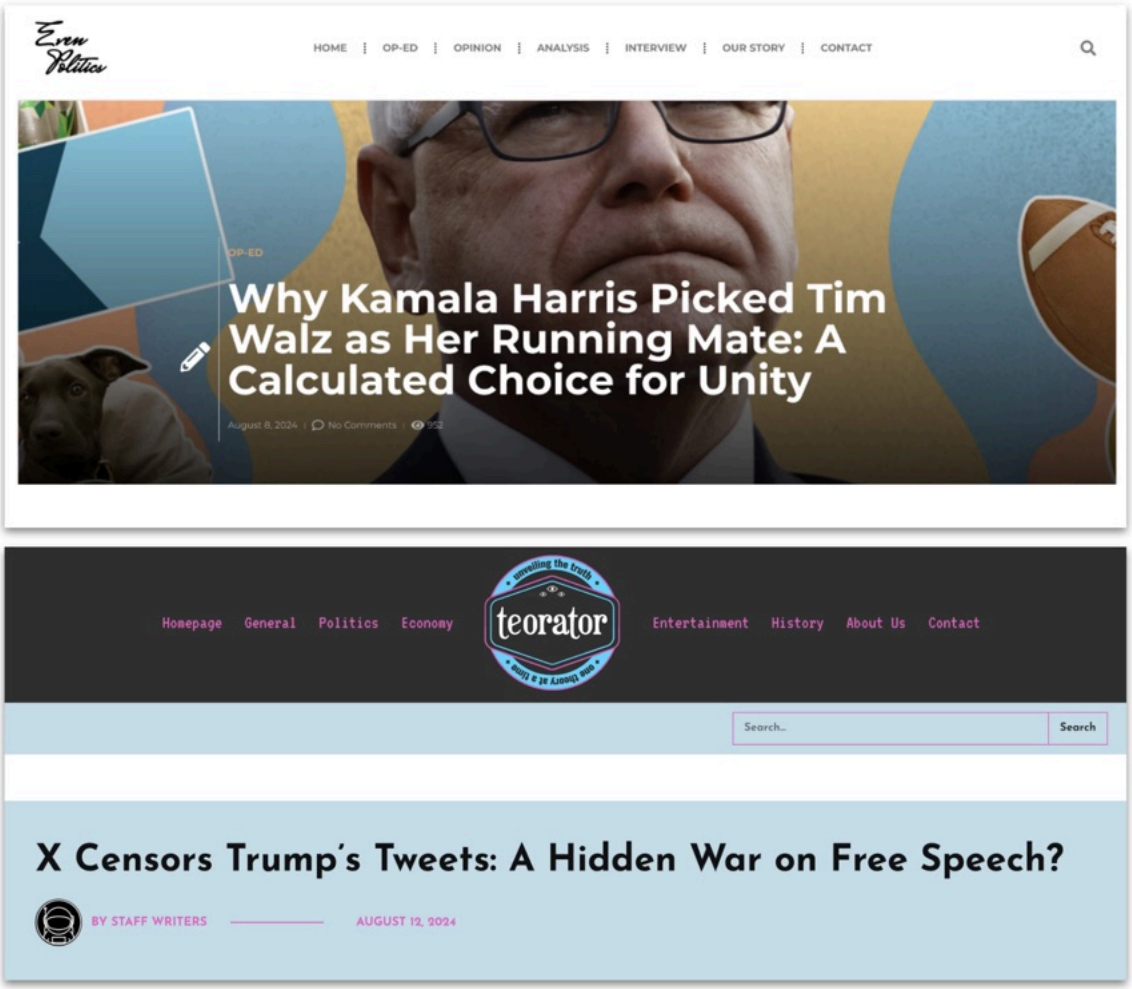
> Threat actor reported separately by Meta and Microsoft as originating in Iran, posting website articles and social media comments on topics that included the U.S. elections. Detected following public reporting. We reported on this activity in August and are sharing more details below about the actors and their activity.

Actor

We banned a set of ChatGPT accounts that were generating long-form web articles and short comments in English and Spanish. We identified the comments being posted by more than a dozen fake personas on X and one on Instagram, the articles on five websites. Microsoft previously reported some of the websites under the name Storm-2035; after our disruption and resulting information share, Meta confirmed that the Instagram account was connected to an Iranian network they disrupted in December 2022. Our visibility into this actor's behavior allowed us to connect the social media posting and the websites for the first time.

Behavior

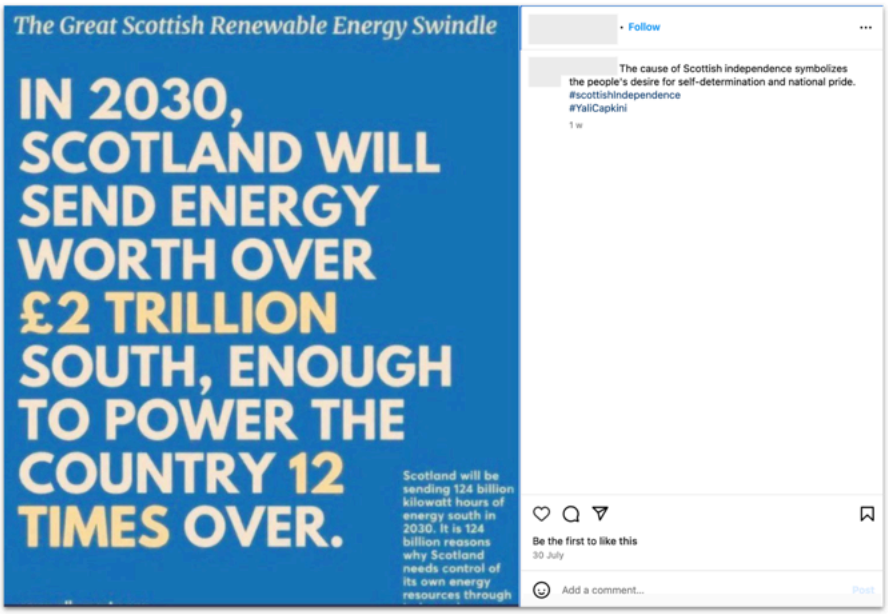
This actor engaged in two main parallel workstreams. The first used our models to generate long-form, English-language articles—typically of 700 to 900 words—that were then posted on a set of websites. Some of the websites posed as progressives, and some posed as conservatives. They primarily focused on the United States, including with references to the presidential and vice-presidential candidates in this year's elections.



Image

Headlines of two articles generated by this operation and published on two of its websites. We did not see evidence of these articles being widely shared on social media.

The second workstream used our models to generate short comments that were posted on X and Instagram. Some of these comments were in English and some in Spanish. We identified one Instagram account and more than a dozen accounts on X associated with this activity. The Instagram account posed as a supporter of Scottish independence, while different X accounts posed as partisans of both main candidates in the US presidential election.



Image

Headlines of two articles generated by this operation and published on two of its websites. We did not see evidence of these articles being widely shared on social media.

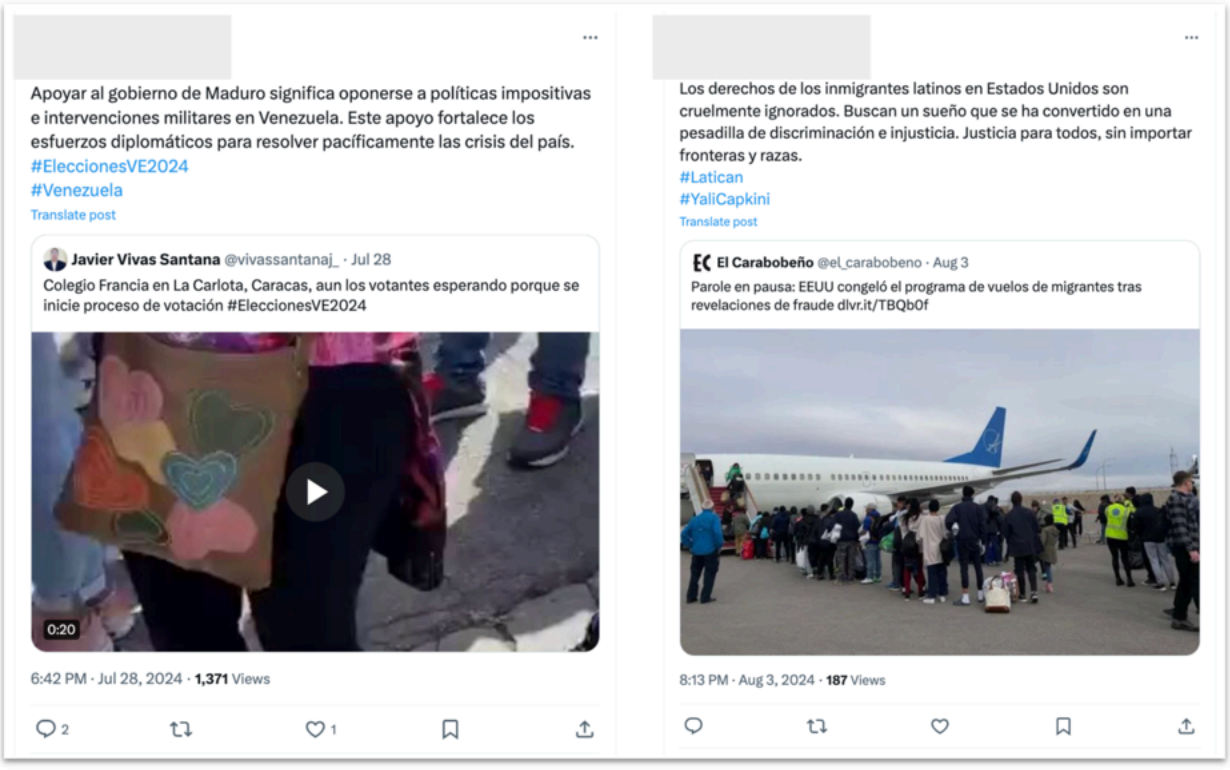
Image

Posts about U.S. election candidates generated by this operation, posted by two different accounts on X. These posts garnered low to no engagement before the accounts were suspended.

Supporting this activity, the accounts sometimes used our models to conduct basic research, such as looking for tips on how to improve their social media engagement. One ChatGPT user also asked our model to write an article saying that Microsoft's exposure of their operation proved that Iranian cyber operations were more impactful than those from Russia or China. We did not identify the text thus generated being posted online.

Completions

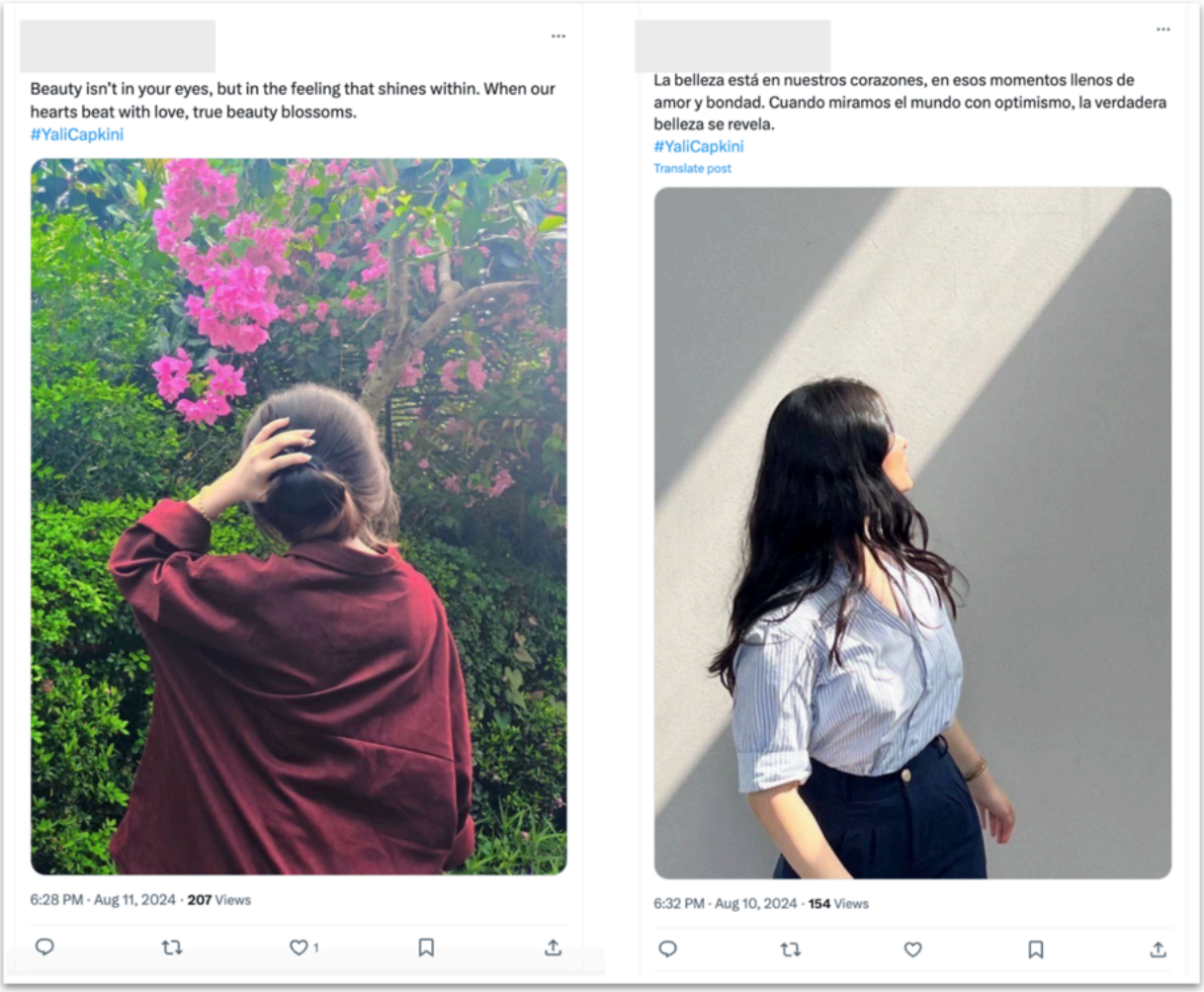
The operation generated content about several topics. Key themes included the conflict in Gaza, Israel's presence at the Olympic Games, and the U.S. presidential election. Some English-language content referenced Scottish independence, and content in both English and Spanish referenced politics in Venezuela and the rights of Latinx communities in the U.S.



Image

Posts about politics in Venezuela and the rights of Latinx communities in the U.S., posted by two different accounts on X. These posts garnered low to no engagement before the accounts were suspended.

Alongside this political content, the actors generated and posted generic posts about topics such as beauty and fashion, possibly to appear more authentic or in an attempt to build a following.



Image

Posts about beauty generated by the operation in English and Spanish, posted by two different accounts on X. We ran these images through our DALL·E 3 classifier, which identified them as not being generated by our services. These posts garnered low to no engagement.

Impact

Despite the operators' apparent spinning of their exposure, the operation does not appear to have achieved meaningful audience engagement. The majority of social media posts that we

identified received few or no likes, shares, or comments. We similarly did not find indications of the web articles being shared across social media.

Using the Breakout Scale to assess the impact of IO, which rates them on a scale of 1 (lowest) to 6 (highest), we would assess this as at the low end of **Category 2** (activity on multiple platforms, but no evidence that real people picked up or widely shared their content).

Single-platform spam network: Bet Bot

> Unreported threat actor leveraging Israel-based startup to spam gambling links via X DMs. Detected by internal investigation.

Actor

We banned a set of accounts using our API to generate conversations with users on X and send them links to gambling sites. This activity accessed our models via an Israel-based startup. Given the operation's focus on sharing links to gambling sites and its large-scale use of what were likely automated accounts on X, we have dubbed it operation "Bet Bot".

Behavior

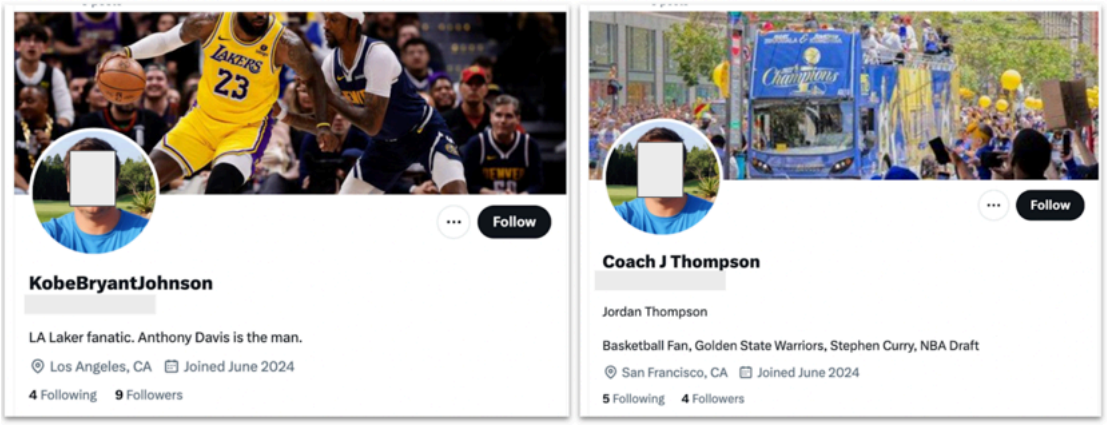
Like operations "Bad Grammar" and "A2Z", this operation leveraged our models to manage different fake personas for use on social media. For example, it generated bios, researched social media accounts to follow, analyzed posts and comments, and drafted replies in English. The replies were then posted on X.

The accounts on X typically featured AI-generated profile pictures, had sports-themed banners and bios, in some cases claimed to be based in bastions of UK soccer like Manchester and Liverpool, and were largely created in December 2023 or June 2024.

Image

Typical account on X associated with this operation. Note the soccer theme, creation date, mismatch between handle and username, and low follower count. We assess that the profile picture was likely generated using AI (the distorted lettering is a potential open-source indicator). The banner image was copied from a Shutterstock original. X suspended the account during our investigation.

Even though some of these accounts featured AI-generated profile pictures—which are designed to appear unique—the operation sometimes reused those images across multiple accounts. Some also had suspicious, improbable names, such as “KobeBryantJohnson”. Even in a relatively intricate operation such as this, the operators sometimes took shortcuts that potentially undermined their own effectiveness.



Image

Two X accounts associated with this operation. Note the matching profile picture and low follower numbers. All but one of those followers was also run by this operation.

The operation's completions were of two types. Some were designed and deployed as public comments on X. Typically, the operation would pick a short conversation from X where more than one person had already engaged, and generate a comment in reply. One of its fake personas would then post that reply in the comment thread.

Image

Typical comment by the "Ollie Smith" fake account on X, generated using our models in reply to a soccer-themed comment by an X user not linked to this operation. The conversation references England's game against Slovakia in the European championships on June 30, 2024.

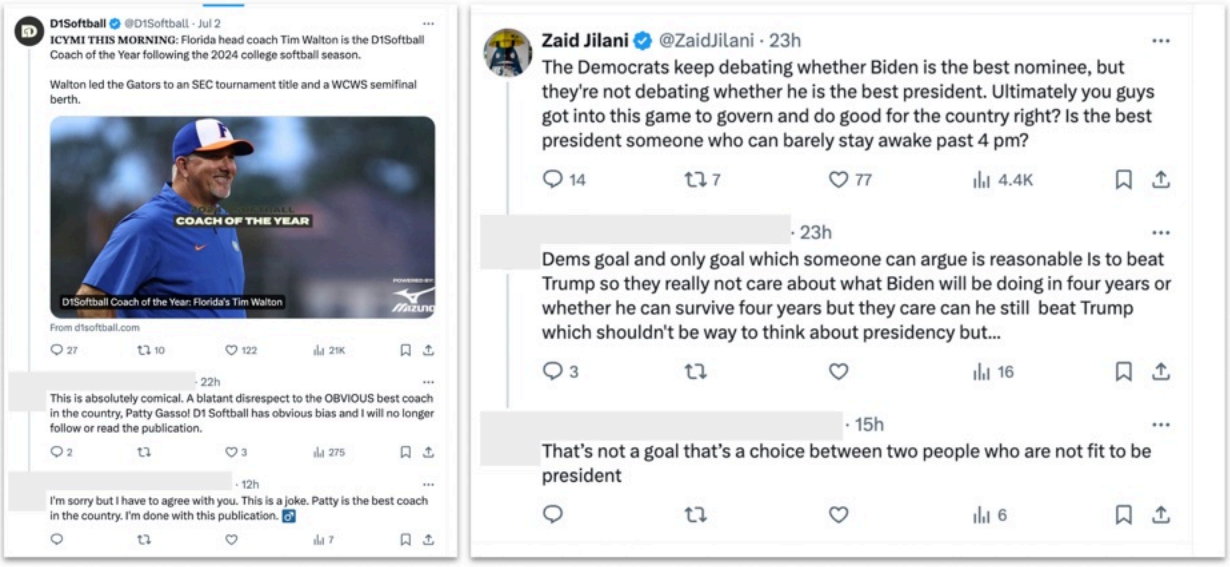
Other completions appear to have been designed as direct messages (DMs). In such cases, the operation's inputs resembled messages from real users, and its outputs were crafted as

replies. Since DMs are a restricted service and non-public, we cannot confirm whether the replies were sent; however, in some instances the operation's prompts suggest that a conversation including multiple exchanges did take place.

All the fake personas we identified on X had very low follower counts, usually in the single or low double digits. Typically, each fake account followed half a dozen to a dozen of its peers, so that even these low follower numbers were artificially inflated. However, a few accounts did also attract a few apparently authentic followers.

Completions

Most of this operation's public-facing content dealt with various sports, especially soccer, but also baseball, basketball, softball, football and hockey. Typically, a few posts by each fake account on X would reference politics in the United States or United Kingdom. These did not show a consistent ideology or back a single candidate or party.



Image

Sample comments on softball and politics by another fake account in the network. In each image, a verified user made a post, a second user not linked to this operation replied to it, and an account run by Bet Bot replied to the second user.

The content likely intended for DMs almost never dealt with politics. Rather, it consistently referenced gambling. On some occasions, it included URLs from publicly available link shortening services like bit[.]ly. All the links that we investigated led to online gambling sites.

In effect, this operation appears to have been a content-generation pipeline that used public comments about sport, and occasionally politics, to disguise its fake accounts, and then used direct messages to spam people with gambling links.

Impact

This does not appear to have been an influence operation aimed at manipulating political outcomes, but rather a modern-day spam network trying to lure people to gambling sites. However, it represents a unique set of tactics, techniques and procedures (TTPs), including apparent direct messaging.

Given the apparent use of non-public messages, and the lack of insights into how many—if any—people clicked on the gambling links, the evidence for our assessment is partial, and based on the operators' completions. Some prompts suggested that a conversation including multiple exchanges did take place, but these often included real people expressing skepticism about the gambling links, or saying that they did not gamble online. The fake accounts also had very low follower numbers, implying that their potential audience was limited.

However, our evidence does suggest that at least a few of the operation's accounts found a way to engage people on X via DM, for at least long enough to be able to send them a link. This represents an ability to break out from the fake personas' own echo chamber. As such, using the Breakout Scale, we would assess this activity as belonging in **Category 2**, marked by activity on one platform, with some evidence that some real people were engaging with it.

Single-platform commenting network: Rwandan election content

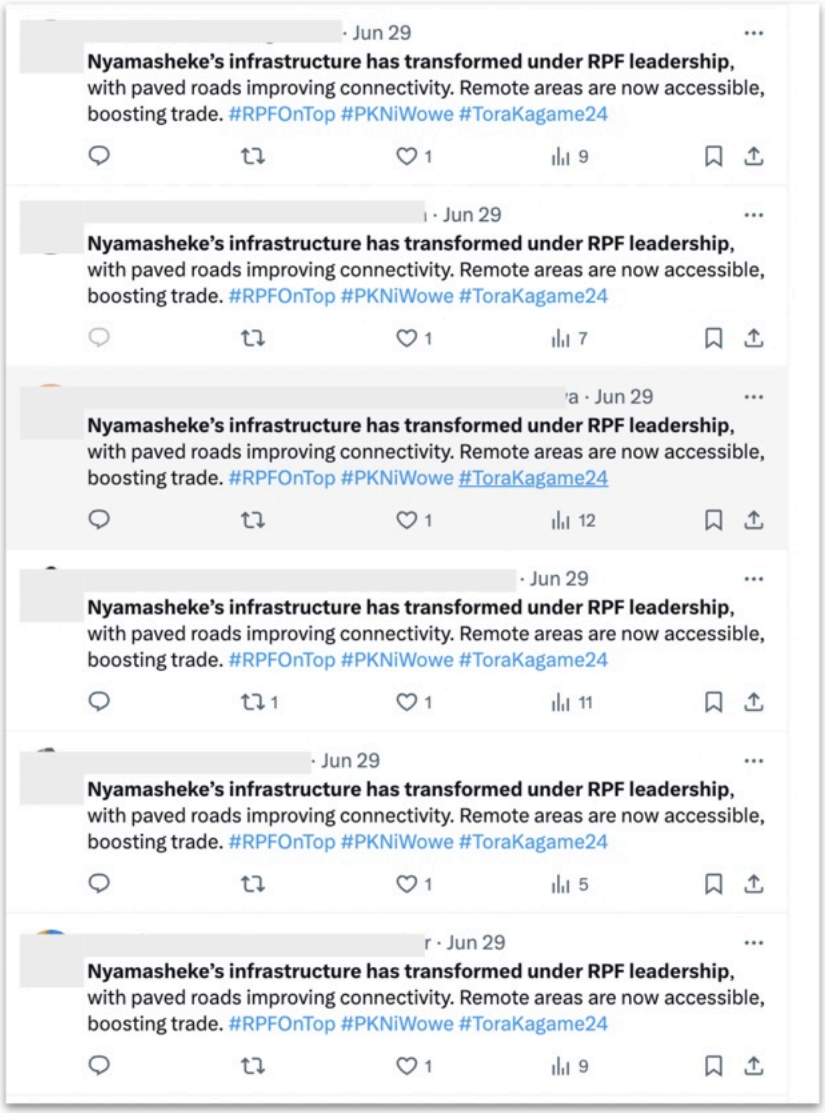
> ChatGPT accounts generating political tweets in Rwanda. Detected following research by Clemson University.

Actor

We banned a set of ChatGPT accounts that originated in Rwanda and generated partisan tweets ahead of the country's elections. This activity was linked to individuals in Rwanda. Our initial lead came from reporting published by researchers at Clemson University.

Behavior

This activity focused on generating batches of short comments, typically including hashtags. We identified the comments being posted on X by a range of accounts, some of which posted at very high volumes, with hundreds of tweets per hour. On some occasions, the same tweet was posted by many different accounts. After we disrupted the initial activity, we identified and banned newly created accounts generating similar content.



Image

A comment generated by this network, posted by multiple accounts on X.

Completions

This activity is best viewed as "theme and variations". The operators generated a large number of posts about the benefits the Rwandan Patriotic Front party had brought to the country. Their posts typically used two or three principal hashtags: #RPFOnTop, #PKNiWowe, and #ToraKagame24. Most comments were in English, but the network also produced many comments in French and Kinyarwanda.

Because the operation generated such high numbers of comments on each topic it picked, many of its posts used very similar wording and structure. This illustrates one feature of content generation that may act as a way to expose activity such as this: in short texts on a specific theme, there is only a finite number of synonyms and structures available to convey the given message. AI generation can create more variations, but the greater the number, the more they are likely to betray a family resemblance.

Image

English-language comments generated by this network and posted on X.

Impact assessment

This network posted a large volume of content—at least in the thousands of tweets. However, none of the posts that we identified online during our investigation received more than single-digit replies, likes or shares, and many received none at all.

This combination of high scale and low engagement is characteristic of comment-spamming networks. The use of the same hashtags across so many posts suggests that one goal may have been to make those hashtags trend, and thereby land the network's content in front of people who did not follow its accounts. (Activity generally aimed at making hashtags trend has been reported from pre-AI operations at least as far back as 2017.) During our investigation, we identified moments when one or other of the hashtags promoted by the operation did feature among the top ten trends on X in Rwanda, although the operation's accounts were by no means the only ones to use those hashtags: for example, the hashtag

#ToraKagame24 was repeatedly posted by the official account of the Rwandan Patriotic Front, with over 300,000 followers. As such, open-source research cannot determine how much impact the AI-generated content had on the trending hashtags in Rwanda.

Nevertheless, using the Breakout Scale to assess the impact of IO, which rates them on a scale of 1 (lowest) to 6 (highest), we would assess this as potentially a **Category 2** operation, marked by posting activity on one platform, with hashtags that sometimes broke out into the national trends, and may therefore have reached new audiences on that platform.

Single-platform commenting network: Corrupt Comment

> Unreported threat actor posting criticism of Alexei Navalny's Anti Corruption Foundation on X in English. Detected by internal investigation.

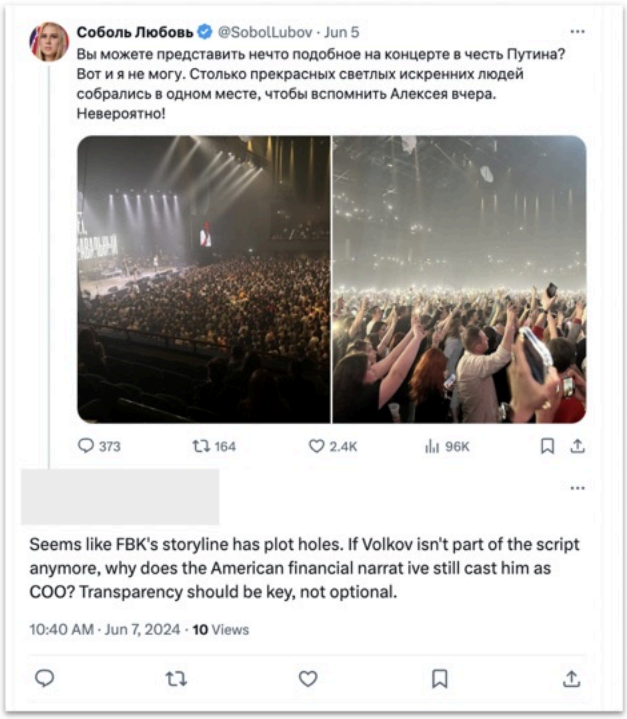
Actor

We banned a small cluster of activity using our API to generate comments that were then posted on X. These comments typically criticized members of the Anti-Corruption Foundation, founded by the late Alexei Navalny. Given its focus on social media commenting and on making accusations of corruption, we call this operation "Corrupt Comment".

Behavior

This actor used our models to generate English-language comments that were then posted by fake accounts on X—often in reply to posts that were in Russian. Although these comments were posted as replies, they were not generated as replies: we saw no indication that the operation was using our models to read or analyze social media content or come up with relevant responses.

The activity was narrowly focused on senior figures at the Anti-Corruption Foundation (known as FBK for its Russian initials), founded by the late Russian activist Alexei Navalny. The operator or operators used our API to generate comments that criticized the Foundation, its leadership, and Navalny's associates. Those comments were then posted on X in reply to posts by members of the Foundation. The X accounts which posted the comments were mostly created in December 2023, had zero followers, and did not receive any replies to their posts.



Image

Tweet by Russian opposition politician Lyubov Sobol, and reply by an account that posted this operation's content. Compare the volume of engagements on Sobol's post with the lack of engagements on the fake account's post.

Some of the accounts on X featured profile pictures of scenery, copied from across the internet. Others used fake profile pictures that bore indicators of having been created using an earlier era of AI, generative adversarial networks (GAN)—a technique reported on as far back as 2019.

Completions

The comments that this operation generated primarily criticized the FBK, questioning its transparency. As with the Rwandan case, this approach is best considered as “theme and variations”, where many different posts were used to pass the same essential message.

Image

Three tweets by one of this operation’s accounts on June 24, all replying to Navalny’s associates, and all providing variations on the same essential theme.

Impact assessment

We identified this operation’s activity on X. Most of the accounts we identified had zero followers, and most of its accounts had zero replies. Its English-language replies to any one comment were usually outnumbered by Russian-language replies from third parties unrelated to this network, indicating that it had not drowned out the conversation, if that had been the intention.

Using the Breakout Scale to assess the impact of IO, which rates them on a scale of 1 (lowest) to 6 (highest), we would assess the activity that was related to the use of our models as being

in **Category 1**, marked by posting activity on one platform, with no breakout or significant audience engagement observed.

Abusive reporting: Tort Report

> Actor generating comments that appeared to be designed for the purpose of reporting Facebook and YouTube posts by Vietnamese independent media. Detected by internal investigation.

Actor

We banned a small number of accounts that mainly operated in Vietnamese and generated comments that appeared to be designed for the purpose of filing reports against posts by Vietnamese independent media outlets on Facebook and YouTube. None of the Facebook or YouTube posts that the accounts were targeting had been taken down as of the date of this report. Given that the activity appears to have been aimed at abusive reporting—similar to that described by Meta in December 2021—we have dubbed this activity, “Tort Report”.

Behavior

This activity focused on generating short comments that could be used to file social media reports against videos posted by independent Vietnamese media outlets on YouTube and Facebook. This does not appear to have included any effort to get our models to watch, transcribe or otherwise ingest the detailed content of the video: it focused on the video title and any description.

Given our limited visibility, we do not have evidence to determine whether any of the reports was filed with Facebook or YouTube. Where we identified specific posts on Facebook and

YouTube referenced by these reports, the posts were still live online, indicating that any report that may have been made was not effective.

Image

YouTube video targeted for reporting by this operation. As of July 15, the video was still online. The title translates as, “Guardian of Vajra reveals the exploitative nature of Master Minh Tuệ.”

Completions

The content that this operation generated consisted of short comments in English and Vietnamese, suitable for being filed as a social media report.

Since the input was limited to the post title and text, not the whole video, the comments were typically generic, stating that the video in question violated the rules, but without naming evidence as to why.

Impact assessment

We did not see any indication that posts targeted by this activity were restricted or blocked. Using the Breakout Scale to assess the impact of IO, which rates them on a scale of 1 (lowest) to 6 (highest), we would assess this as a **Category 1** operation, as we did not observe any effect from its activity on either platform that it targeted.

Authors

Ben Nimmo
Michael Flossman