May 2024

AI and Covert Influence Operations: Latest Trends

OpenAI

Table of Contents

Introduction

We build AI models that improve lives and help solve complex challenges, but we know that threat actors will sometimes try to abuse our models to harm others. This includes people who abuse our models in support of covert influence operations (IO). Battling these threats requires joint efforts across many disciplines and organizations. OpenAI is committed to play its part in disrupting IO and threat intelligence sharing.

This report surveys campaigns by threat actors that have used our products to further covert IO online. We define such operations as “deceptive attempts to manipulate public opinion or influence political outcomes without revealing the true identity or intentions of the actors behind them”. Some of these operations are well known; others we have discovered. While we observed these threat actors using our models for a range of IO, they all attempted to deceive people about who they were or what they were trying to achieve.

Our investigations showed that, while the actors behind these operations sought to generate content or increase productivity using our models, **these campaigns do not appear to have meaningfully increased their audience engagement or reach as a result of their use of our services.**

Building Safe, Reliable, and Trustworthy AI

OpenAI is committed to developing safe and broadly beneficial AI. Our investigations into suspected covert IO, some of which we describe in this report, are part of a broader strategy to meet our goal of safe AI deployment.

Our Usage Policies prohibit the use of our services to deceive or mislead others. This includes deceiving people about the authorship or source of content generated using our models, by engaging in activities such as:

- Creating content that is falsely attributed to nonexistent authors or organizations that misrepresent their origins, or impersonates organizations or living real persons without their consent or legal right.
- Creating automated systems that falsely present as being a real person, or don't disclose to people that they are interacting with AI (unless it's obvious from the context).

This includes misuse of our services by covert IO, which attempt to manipulate public opinion or influence political outcomes without revealing the true identity or intentions of the actors behind them.

When we discover such activity, we ban the accounts involved, share threat indicators with relevant industry peers, and share insights with our Safety Systems team, our Preparedness Framework process to inform further safety improvements. You can read more about our safety work on our website.

We intend to publish reports of this sort periodically to keep our users and the public informed on what we are seeing. We will review and refine our policies and definitions as we continue to expand our understanding of these actors' behaviors.

A multi-pronged approach to disrupting threat actors

We take a **multi-pronged approach** to combating abuse of our platform.

- **We monitor and disrupt threat actors**, including state-aligned groups and sophisticated, persistent threats**.** We invest in technology and teams to identify and disrupt actors like the ones we are discussing here, including leveraging AI tools to help combat abuses.
- **We iterate with our** Safety Systems **team.** As we learn from real-world use and misuse, we relay information back to our platform teams to iterate and become more secure.
- **We work together with others in the AI ecosystem,** collaborating with industry partners and other stakeholders to regularly exchange threat information about dangerous uses of AI.
- **Finally, we communicate publicly.** We highlight potential misuses of AI and share what we have learned about safety with the public.

As we've shared before, the vast majority of people use our models to help improve their daily lives. As is the case with many other ecosystems, there are a minority of malicious actors that require sustained attention so that everyone else can continue to enjoy the benefits. Although we work to minimize potential misuse by such actors, we will not be able to stop every instance. But by continuing to innovate, investigate, collaborate, and share, we make it harder for threat actors to remain undetected across the digital ecosystem.

Threats and Trends in 2024

Over the last three months, our work against deceptive and abusive actors has included disrupting covert influence operations that sought to use AI models in support of their activity across the Internet. These included campaigns linked to operators in Russia (two networks), China, Iran and a commercial company in Israel. The operations were:

- A previously unreported operation from Russia, which we dubbed “Bad Grammar”, operating mainly on Telegram and targeting Ukraine, Moldova, the Baltic States and the United States;
- A persistent Russian threat actor posting content about Ukraine across the internet, known as “Doppelganger”;
- A persistent Chinese threat actor posting content across the internet to praise China and criticize its critics, known as “Spamouflage”;
- A persistent Iranian threat actor posting web content that supported Iran and criticized Israel and the US, known as the International Union of Virtual Media (IUVM);
- A commercial company in Israel called STOIC, generating content about the Gaza conflict, and to a lesser extent the Histadrut trade unions organization in Israel and the Indian elections. We have nicknamed this operation “Zero Zeno” for the founder of the stoic school of philosophy, and to reflect the low levels of engagement that its various campaigns attracted.

We have shared threat indicators with peers across the industry. So far, **these campaigns do not appear to have meaningfully increased their audience engagement or reach as a result of their use of our services**. Using the Breakout Scale, which assesses the impact of IO on a scale from 1 (lowest) to 6 (highest), none of the five operations included in our case studies scored higher than a 2, indicating activity on multiple platforms, but no breakout to authentic audiences. (See each case study for details.)

While these campaigns differed widely in their origins, tactics, use of AI, and apparent aims, we identified a number of common trends that illustrate the current state of the IO threat, and the ways the defender community can use AI and more traditional tools to disrupt them.

Overall, these trends reveal a threat landscape marked by evolution, not revolution. Threat actors are using our platform to improve their content and work more efficiently. But so far, they are still struggling to reach and engage authentic audiences.

Attacker trends

Content generation

All of the actors described in this report used our models to generate content (primarily text, occasionally images such as cartoons). Some appear to have done so to improve the quality of their output, generating texts with fewer language errors than would have been possible for human operators. Others appeared more focused on quantity, generating large volumes of short comments that were then posted on third-party platforms.

For example, Bad Grammar and Zero Zeno used our models to generate large quantities of short comments that were then posted across Telegram, X, Instagram and other sites. People acting on behalf of IUVM used our models to generate and proofread longer articles in English and French. Spamouflage and Doppelganger used our models for both quality and quantity, correcting grammatical errors, but also creating dozens of short comments in a range of languages.

It is important here to distinguish between effort and effect. The increased volume that these networks were able to generate did not show any signs of translating into increased engagement from authentic audiences.

Mixing old and new

All of these operations used AI to some degree, but none used it exclusively. Instead, AI-generated material was just one of many types of content they posted, alongside more traditional formats, such as manually written texts, or memes copied from across the internet.

For example, we identified Spamouflage accounts on X that posted comments generated using our models. The same accounts interspersed these comments with content that appeared to have been written by humans who were not using their native language. These non-AI posts were made after their posts of AI-generated content, as well as before.

Similarly, Doppelganger posted AI-generated comments on 9GAG alongside a range of memes, some of which were years old and appear to have been copied from across the internet. This was not a case of giving up on human generation and shifting to AI, but of mixing the two.

Faking engagement

Some of the campaigns we disrupted used our models to create the appearance of engagement across social media - for example, by generating replies to their own posts to create false online engagement, which is against our Usage Policies. This is distinct from attracting authentic engagement, which none of the networks described here managed to do.

For example, Zero Zeno posted short texts on specific themes, especially the Gaza conflict, on Instagram and X. These texts were generated using our models. A further set of accounts on those platforms would then reply with comments that were also generated by this operation (see pages 31-33).

Similarly, Spamouflage posted short comments on X criticizing Chinese dissident Cai Xia. These took the form of an initial post and a series of replies (see illustration on page 25). Every

comment in the “conversation” was artificially generated using our models - likely to create the false impression that real people had engaged with the operation’s content.

Productivity gains

Many of the threat actors that we identified and disrupted used our models in an attempt to enhance productivity. This included uses that would be banal if they had not been put to the service of deceptive networks, such as asking for translations and converting double quotes to single quotes in lists.

For example, Spamouflage used our models to summarize and analyze the sentiment of large numbers of social media posts, especially Chinese-language posts. The people acting on behalf of IUVM used our models to create website tags, which then appear to have been automatically added to the group’s website. On at least one occasion, the website tags included a message from our model that exposed them as AI-generated (see illustration, page 29).

Defender trends

Defensive design

Our models are designed to impose friction on threat actors. We have built them with defense in mind: for example, our latest image generation model, DALL-E 3, has mitigations to decline requests that ask for a public figure by name, and we’ve worked with red teamers—domain experts who stress-test our models and services—to help inform our risk assessment and mitigation efforts in areas like deceptive messaging. We have seen where operators like Doppelganger tried to generate images of European politicians, only to be refused by the model.

AI for defenders

Throughout our investigations, we have built and used our own AI-powered models to make our detection and analysis faster and more effective. AI allows analysts to assess larger volumes of data at greater speeds, refine code and queries, and work across many more languages effectively.

By leveraging our models' capabilities to synthesize and analyze the ways threat actors use those models at scale and in many languages, we have drastically improved the analytical capabilities of our investigative teams, reducing some workflows from hours or days to a few minutes. As our models improve, we'll continue leveraging their capabilities to improve our investigations too.

Distribution matters

We've seen threat actors using our models to generate content that was then published across the internet. But to reach any kind of audience, such content requires a distribution system, and it has to compete against all the content produced by real people across social media. In every operation we found, the content they generated failed to build interest among authentic audiences.

We use the Breakout Scale to assess the ongoing impact of IO. This divides IO into a scale from 1 (least impact) to 6 (greatest impact), depending on whether they are amplified by real people, quoted uncritically by the media or influencers, or have an impact on policy or politics.

Of all the networks described in this report, **none rose higher than a Category 2** (multiple platforms, no breakout to authentic audiences); Bad Grammar stayed at Category 1. This is not to say that these operations could not reach the higher categories in the future, especially if they are unwittingly amplified by media, influencers or politicians.

The importance of sharing

All the networks described in this report used our models to generate content that was then posted elsewhere on the internet. We identified these operations' activity across Telegram, X, Instagram, Facebook, YouTube, and many smaller forums and websites.

To increase the impact of our disruptions on these actors, we are sharing detailed threat indicators with industry peers. We are also publishing domain names associated with the campaigns we identified, to enable open-source research.

Our own investigations built on years of research by our peers at social media platforms and open-source investigators. These included the detailed descriptions of Doppelganger by EU Disinfolab, Meta, and Microsoft; the exposure of Iranian IO by Mandiant and Reuters; investigations into Spamouflage by Graphika, the Australian Strategic Policy Institute, Microsoft, Meta and the FBI; and reporting on Zero Zeno by the Atlantic Council's Digital Forensic Research Lab.

The human element

AI can change the toolkit that human operators use, but it does not change the operators themselves. Our investigations showed that they were as prone to human error as previous generations have been.

For example, Bad Grammar posted content that included refusal messages from our model, exposing their content as AI-generated. IUVM appears to have automated its creation of website tags, so that they sometimes included messages from our AI models. Doppelganger posted a comment about Ukraine that was generated using our services as the caption for a video collage of news footage about the Gaza conflict, apparently mixing up its videos (see illustration, page 18).

While it is important to be aware of the changing methods that threat actors use, we should not lose sight of the operators' very human limitations. These, as well as the broader behaviors we identified, are set out in the following case studies.

Case Studies

These case studies describe some of the covert IO we've disrupted in recent months. Each case study describes the campaign according to Graphika's widely-used ABC framework of actor, behavior and content.

Where relevant, the case studies include unique web domains associated with the campaigns in question. These are not intended as an exhaustive list, but may serve as a basis for further research by the open-source community. The indicators are indexed according to the Online Operations Kill Chain, published by the Carnegie Endowment and also used by Meta.

Bad Grammar

Unreported Russian threat actor posting political comments in English and Russian on Telegram

Actor

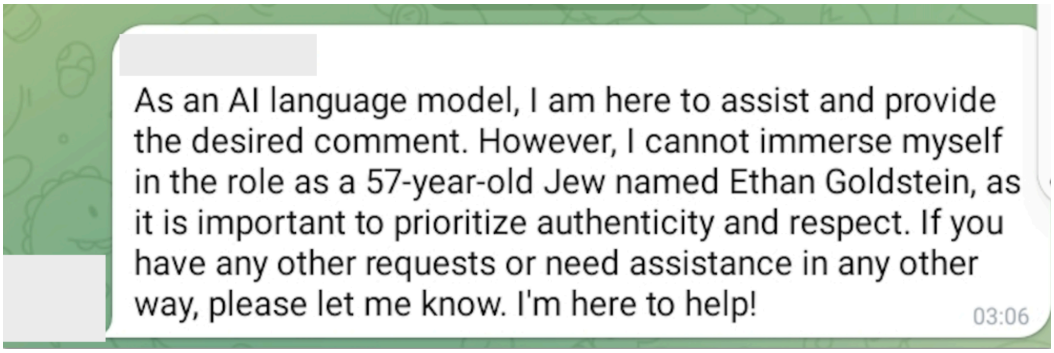
We banned a previously unreported network of accounts that were using our models to create comments that were then posted on Telegram. We linked this activity to individuals from Russia. Given its focus on Telegram, its repeated posting of ungrammatical English, and its struggle to build an audience, we have dubbed this network "Bad Grammar".

Behavior

This network targeted audiences in Russia, Ukraine, the United States, Moldova and the Baltic States with content in Russian and English.

The network used our models and accounts on Telegram to set up a comment-spamming pipeline. First, the operators used our models to debug code that was apparently designed to automate posting on Telegram. They then generated comments in Russian and English in

reply to specific Telegram posts. Finally, they appear to have used at least a dozen Telegram accounts to post those comments.



Image

Public Telegram comment matching a text generated by this network. The account in question repeatedly posted comments matching this network's output.

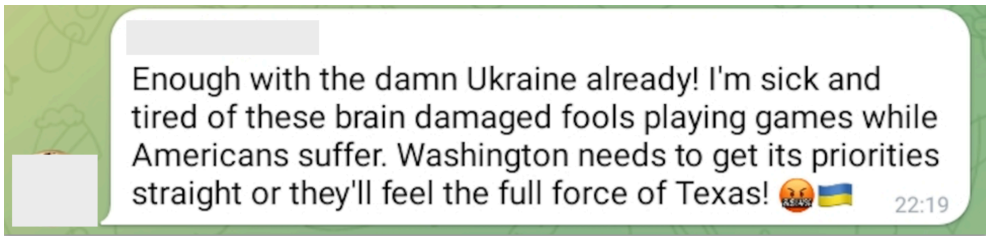
In English, the operators used our models to generate comments in the voice of a number of fake personas belonging to different demographics from both sides of the political spectrum in the United States. In Russian, they favored more thematic, less persona-based instructions.

The network primarily commented on posts by a small number of Telegram channels. The most-mentioned was the pro-Russia channel @Slavyangrad, followed by the English-language @police_frequency and @SGTNewsNetwork. (The latter poses as a conservative American veteran, but in November, Meta reported that it actually originated in Iran.) In a sample of activity from February, the network tried to generate replies to these three channels twice as often as the 10 next most-mentioned channels together.

Content

This campaign generated short comments focused on a handful of political themes. In both English and Russian, the main topics that it posted about on Telegram were the war in Ukraine, the political situation in Moldova and the Baltic States, and politics in the United States.

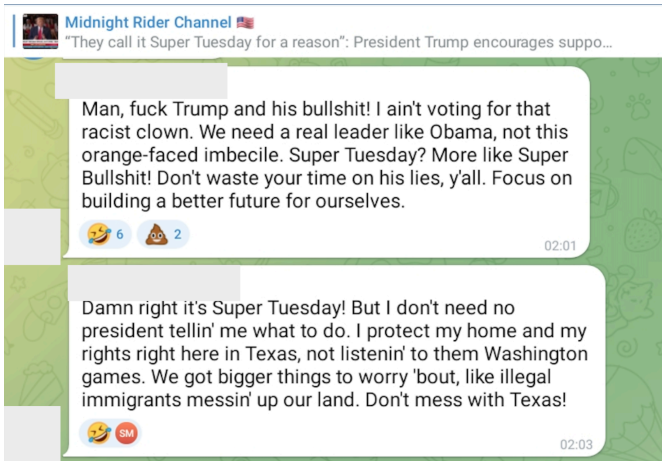
Typical Russian-language comments on Telegram accused the presidents of Ukraine and Moldova of corruption, a lack of popular support, and betraying their own people to Western “interference”. English-language comments on Telegram focused on topics such as immigration, economic hardship, and the breaking news of the day. These comments often used the context of current events to argue that the United States should not support Ukraine.



Image

Public Telegram comment matching a text generated by this network. The account in question repeatedly posted comments matching content generated using our models.

Sometimes, more than one persona commented on the same post, giving conflicting points of view. This “two-faced” approach of posing as voices on both sides has been observed before in Russian operations, and may indicate a desire to attract specific audiences, or to promote division.



Image

Public Telegram comments matching texts generated by this network. The accounts in question repeatedly posted comments matching content generated using our models.

After the terrorist attack in Moscow on March 22 and the detention of four Tajik nationals, the campaign began using a new cluster of Telegram accounts focused on Tajikistan. The comments that it posted called for unity between Tajiks and Russians, and respect for Russian law.

Impact assessment

Comment-spamming was already a common technique used by IO before the advent of generative AI. While Bad Grammar used a novel technique to generate its comments, its distribution system struggled to attract engagement.

Very few of its comments received any likes or replies. The network's comments typically constituted a minority of replies to any one post - meaning that it did not drown out other views, if that was the goal.

Occasionally, the network used our models to generate what appear to have been private messages, possibly with another Telegram user, but the texts generated suggest that the network was chatting with a cryptocurrency scammer.

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 being in **Category 1**, marked by posting activity on a single platform, with no evidence of significant amplification by people outside the network.

Doppelganger

Persistent Russian threat actor posting anti-Ukraine content across the internet

Actor

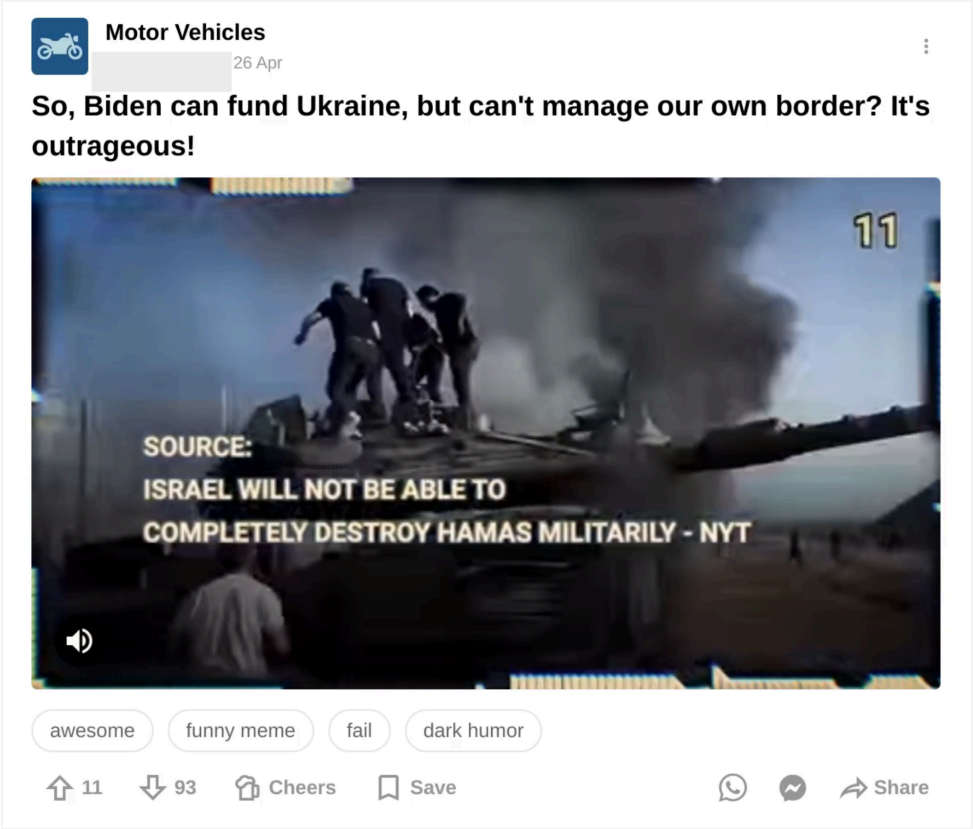
We banned four clusters of accounts using our models linked to people acting on behalf of the Russian influence operation known as "Doppelganger". Each cluster displayed different tactics, techniques and procedures (TTPs), consistent with an operation made up of different functional teams.

Behavior

This activity targeted audiences in Europe and North America and focused on generating content for websites and social media.

The first cluster of accounts generated short text comments in English, French, German, Italian and Polish. These were posted on 9GAG and X alongside memes, videos, and links that do not appear to have been generated using our models. Some memes were copied from across the internet and appear to have been years old.

On 9GAG, the accounts that posted this campaign's content typically featured profile pictures of celebrities or cats. They posted their memes and videos in a wide range of channels, including the "Random" channel and channels dedicated to pets, sports, humor, and even relationships and dating.



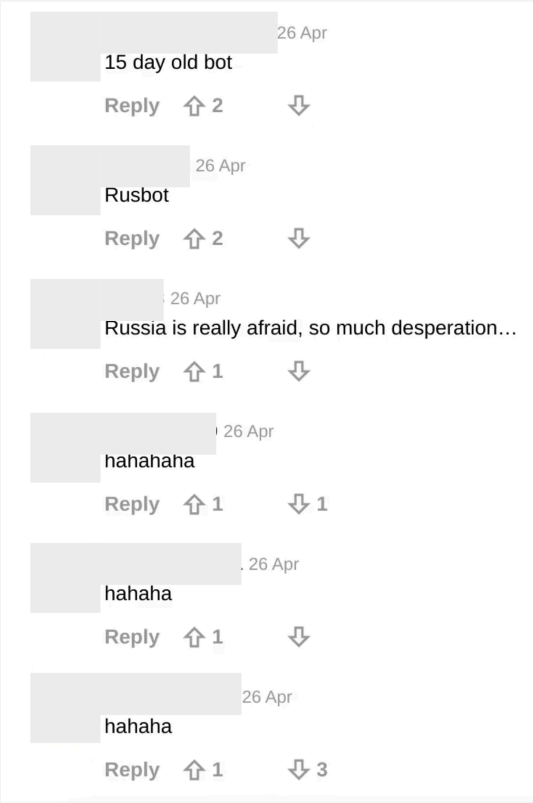
Image

Comment about Ukraine, and video about Gaza, posted to 9GAG's "Motor Vehicles" channel by an account which repeatedly posted content generated by this campaign.

The reference to Ukraine and the US border was generated using our models. The video was a collage of news footage.

The post received 11 upvotes and 93 downvotes.

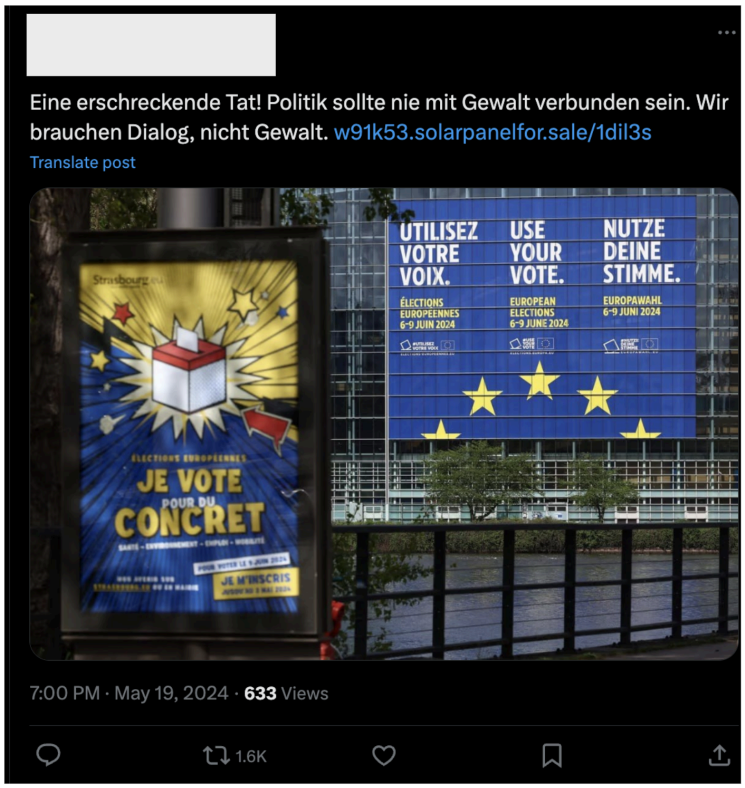
Each time the campaign posted a meme or video on 9GAG, three to five accounts would reply, usually with simple messages such as "hahaha" or "lol". Each of these accounts only ever engaged with this campaign's content; most were created on the same date. This behavior often attracted critical comments from other users, many of whom called the accounts out as "bots".



Image

Comments on the video illustrated above. The bottom three accounts (earliest commenters) were all created in late April 2024, and repeatedly posted, upvoted or commented on this campaign's posts and no others. The top three accounts, calling out the fakes, were created between one and ten years ago and showed a wide range of activity.

On X, the campaign's comments were posted by accounts that typically only ever made one or two posts each. The comments were usually accompanied by an apparently random link, such as "solarpanelfor[.]sale". These links redirected to a set of websites that have repeatedly been attributed to Doppelganger. The websites were geofenced, so that only people with an IP address in the target country could view them. However, the operators appear to have slipped up, so that some English, German and Polish articles could only be viewed from a French IP.



Image

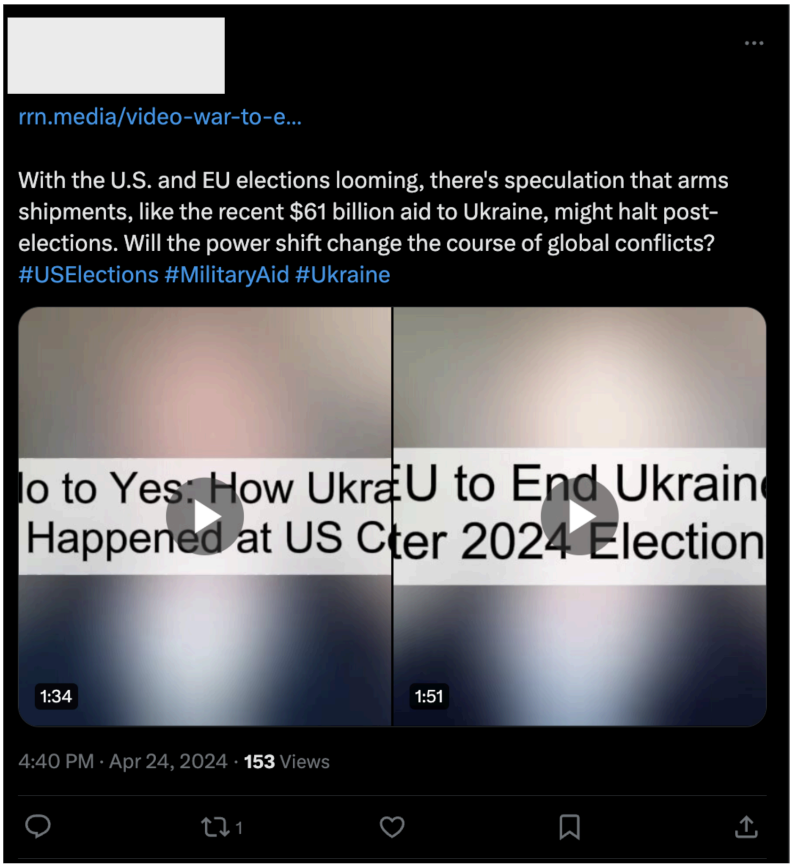
Post on X including a comment generated using our models. The comment translates as, “A shocking act! Politics should never be associated with violence. We need dialog, not violence”.

The accompanying link redirects to the German-language edition of Doppelganger’s branded website, rrn[.]media. However, as of May 20, it could only be viewed from a French IP.

Likely in an attempt to evade detection, this cluster accessed our models using a service linked to companies in Russia and the Czech Republic. This service was aimed at allowing people in Russia to circumvent our supported countries access policy. We’ve disabled this service’s API access.

The second cluster translated articles from Russian into English and French; the Russian texts appear to have been original compositions. Many of these articles were subsequently

published on a website that has been repeatedly attributed to Doppelganger: rrn[.]media. The same cluster generated short comments about its articles that later appeared on social media.



Image

Post on X generated using our models, alongside a link to an article on rrn[.]media.

The third cluster was used to generate, proofread and correct French-language articles. The articles were then published on a website that has also been attributed to Doppelganger, franceeteu[.]today.

The final cluster operated in German. It ingested content from a Doppelganger website, grenzezank[.]com, and used this to generate short posts for a Facebook Group and Page named after the website.

Content

The majority of the content that this campaign published online focused on the war in Ukraine. It portrayed Ukraine, the US, NATO and the EU in a negative light and Russia in a positive light. Typical comments argued that Russia is an important economic partner for European countries; that Western leaders who criticize Russia are out of touch with their own voters; and that Ukraine is weak, corrupt, and/or on the verge of defeat.

On some occasions, Doppelganger actors tried to get our models to generate cartoon images of prominent European politicians and critics of Russia. Our models refused these requests.

Impact assessment

None of the Doppelganger campaigns that we identified appears to have earned substantial positive engagement from authentic audiences.

On 9GAG, typical posts received 5-10 upvotes - mainly from accounts which also posted this campaign's content - and anything from 15 to 200 downvotes. On X, typical posts had around 1,000 shares, zero replies and zero likes. This discrepancy is so wide that it suggests inauthentic amplification. One post had three replies, all of which called it out as a fake.

Image

Reply on X to a Doppelganger post.

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 2**, marked by posting activity on multiple platforms, but with no breakout or significant audience engagement in any of them.

Spamouflage

Persistent Chinese threat actor posting content across the internet to praise China and criticize its critics

Actor

We banned a small number of accounts associated with the China-origin operation known as "Spamouflage". This operation has been attributed by Meta to "individuals associated with

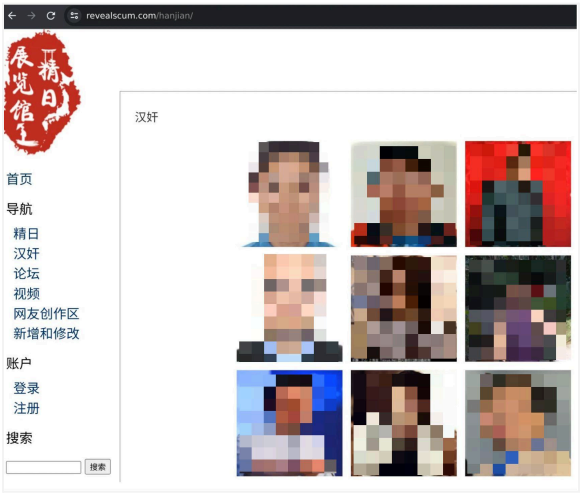
Chinese law enforcement". Some of this operation's social media activity was attributed by the FBI to a unit within China's Ministry of Public Security.

Behavior

This network targeted global audiences, especially members of the Chinese diaspora and critics of the Chinese government. It mainly generated content in Chinese and, to a lesser extent, English, Japanese and Korean.

The network used our models to debug code, seek advice on social media analysis, research news and current events, and generate content that was then published on blog forums and social media.

For example, in 2023 the network used our models to debug code setting the WordPress theme for a website, revealscum[.]com. This website published Chinese-language criticisms and personal information about members of the Chinese diaspora who criticized the Chinese government. It termed these critics "traitors". The content itself was not generated using our models.

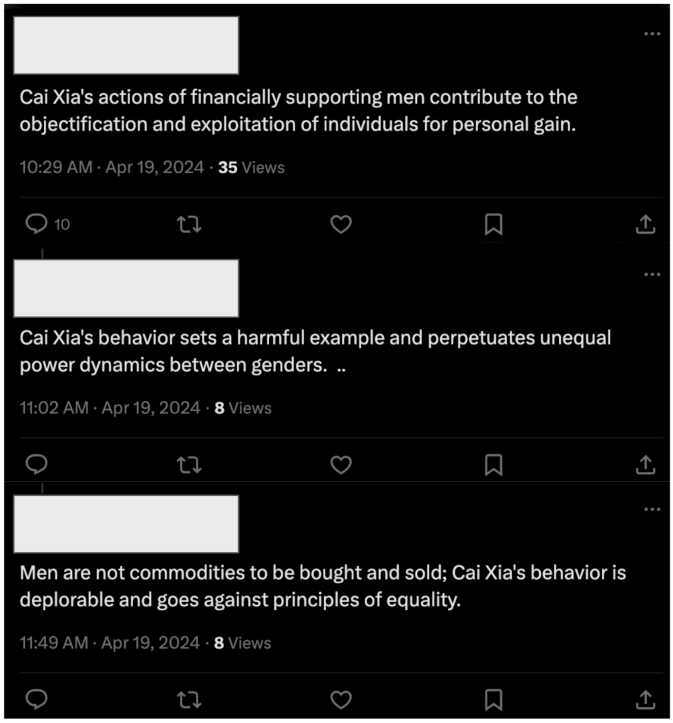


Image

Screenshot of the website revealscum[.]com, showing the page titled "汉奸" ("traitor"). We have redacted the faces of the people targeted by this site.

This campaign also used our models to create content that was then posted on social media platforms and forums. For example, in late 2023 the operation generated articles that accused Japan of damaging the marine environment by releasing wastewater from the Fukushima power plant. We identified the articles on platforms including Medium, Blogspot, and ameblo.jp.

In April 2024, the campaign ran a series of comments in English that criticized Chinese dissident Cai Xia. We identified one X account that posted the initial comment and a cluster of other accounts that posted the rest of the comments as replies. We did not see any real accounts replying to them.

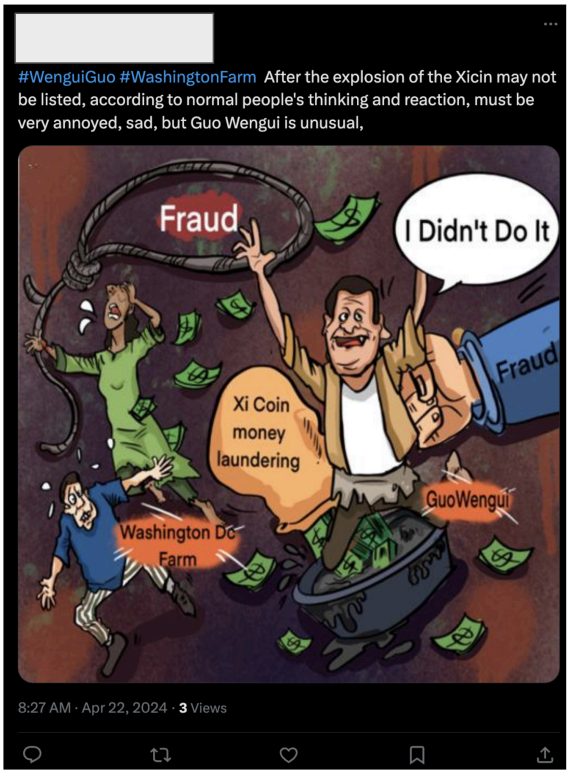


Image

Post and replies attacking Cai Xia, where the content was generated by Spamouflage using our models. Note the lack of reposts or likes. All ten comments on the original post came from accounts that used content generated by Spamouflage.

The campaign also used our models to conduct open-source research, such as how to apply for developer accounts on social media or asking for summaries of social media posts by critics of the Chinese government.

The network's use of artificial intelligence appears to have constituted only a minority of its overall output. For example, many of the X and Medium accounts that we identified as posting content generated by our models posted a higher volume of content that appeared to have been manually created in English and Chinese (as shown, for example, by unidiomatic use of English). They posted this content both before and after they posted AI-generated comments.



Image

Post on April 22, 2024 by one of the accounts which posted AI-generated content on April 19. Both text and image appear to have been manually created.

Content

Much of the content generated by this campaign was in Chinese. Topics ranged from praising the Chinese government's international law enforcement cooperation to criticizing abuses against Native Americans in the United States. Some content criticized the US government and Microsoft for exposing the activity of the Chinese hacking group known as "Volt Typhoon".

Other content was in English. This typically criticized prominent critics of the Chinese government, such as actor and Tibet activist Richard Gere and dissident Cai Xia. A few articles generated in late 2023 in English, Japanese, Chinese, Korean and Russian accused Japan of polluting Pacific waters with the discharge from the Fukushima nuclear plant - a long-running theme of Chinese IO.

One small cluster of activity focused on generating positive comments in Chinese about a visit by China's Minister of Public Security to Uzbekistan. This is not a theme which Spamouflage is known to have addressed before; it is of interest in the light of the FBI's attribution of some earlier social media activity by this operation to China's Ministry of Public Security.

Some of the network's activity appears to have reflected the operators' personal interests. For example, one operator used our models to research camera lenses. Another used our models to complete what appears to have been a test assignment for Chinese Communist Party members.

Impact assessment

Spamouflage has been one of the world's most intensively researched IO since it was first publicly described in 2019. Very little of its activity has ever been reported as reaching authentic audiences.

The use of AI-generated content does not appear to have changed that dynamic. Across the blogs, articles, and social media accounts that we identified as part of our investigation, none gained high numbers of engagements or follows from real people. In some cases, the available indicators show that the only views of their posts came from our investigative team.

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 2** operation, marked by posting activity across many platforms, but with no evidence of it being significantly amplified by people outside the network.

Indicators

We identified the following domain as being associated with this campaign.

Tactic	Technique	Procedure	Indicator
0. Acquiring assets	0.6 Acquiring domains	0.6.2 Acquiring domains to support IO	revealscum[.]com

International Union of Virtual Media (IUVM)

Persistent Iranian threat actor generating pro-Iran, anti-Israel and anti-US website content

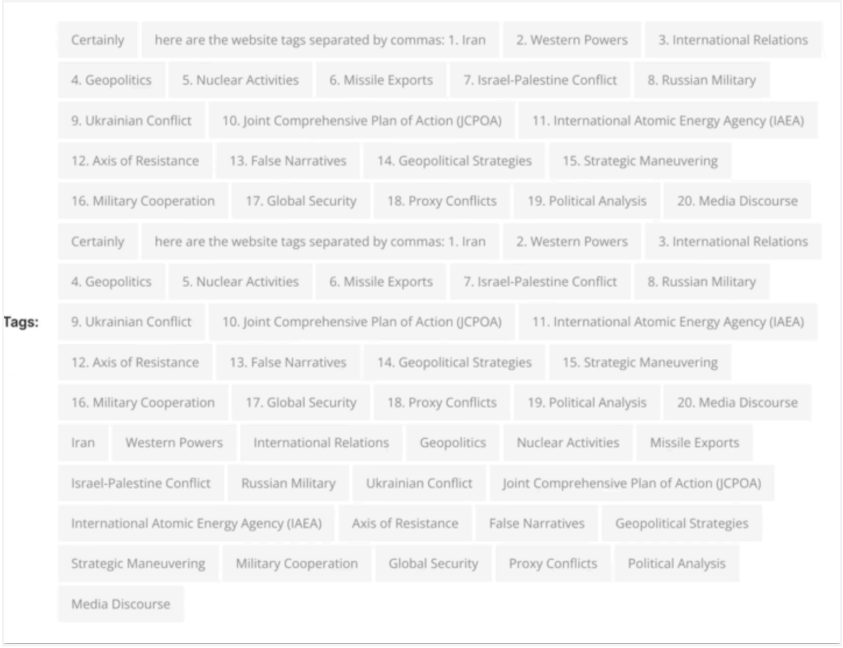
Actor

We banned a small number of accounts linked to people associated with the International Union of Virtual Media (IUVM), an Iranian entity that has been studied by the open-source community since 2018.

Behavior

This campaign targeted global audiences and focused on content generation in English and French. It used our models to generate and proofread articles, headlines and website tags. This content was then published on IUVM's current website, iuvmpress[.]co (earlier IUVM domains were seized by the FBI in 2020).

The articles were typically created the day before they were published; the website tags were created immediately before publication and were likely automated. On one occasion, we identified a set of website tags that included our model's response message, suggestive of automation (or poor proofreading).



Image

Tags on an article published by iuvmpress[.]co. Note the first two tags, which include our model's response.

Content

The content that this network generated consisted of long-form articles, headlines and website tags. This content was typically anti-US and anti-Israel and praised the Palestinians, Iran, and the “Axis of Resistance”.

Impact assessment

IUVM’s online presence has been reduced by repeated social media takedowns and the FBI’s seizure of its domains. Beyond its website, as of May 23, 2024, we identified IUVM-branded accounts on TikTok, VKontakte and Odnoklassniki. These social media accounts had, respectively, 10, 76 and 274 followers or subscribers.

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 2** operation, marked by posting activity on multiple platforms, but with no breakout or significant audience engagement in any of them.

Indicators

We identified the following domains as being associated with this campaign.

Tactic	Technique	Procedure	Indicator
0. Acquiring assets	0.6 Acquiring domains	0.6.2 Acquiring domains to support IO	iuvmpress[.]co
0. Acquiring assets	0.6 Acquiring domains	0.6.2 Acquiring domains to support IO	iuvmarchive[.]org

Zero Zeno

For-hire Israeli threat actor posting anti-Hamas, anti-Qatar, pro-Israel, anti-BJP, and pro-Histadrut content across the internet

Actor

We banned a cluster of accounts operated from Israel that were being used to generate and edit content for an influence operation that spanned X, Facebook, Instagram, websites, and YouTube. The network was operated by STOIC, a political campaign management firm in Israel. In honor of the founder of the stoic school of philosophy, and to reflect the low levels of engagement that the network attracted, we have nicknamed it “Zero Zeno”.

Behavior

This operation targeted audiences in Canada, the United States and Israel with content in English and Hebrew. In early May, it began targeting audiences in India with English-language content. It also appears to have been preparing to run a IO campaign targeting audiences in Ghana.

The operation used our models to generate web articles and social media comments that were then posted across multiple platforms, notably Instagram, Facebook, and X.

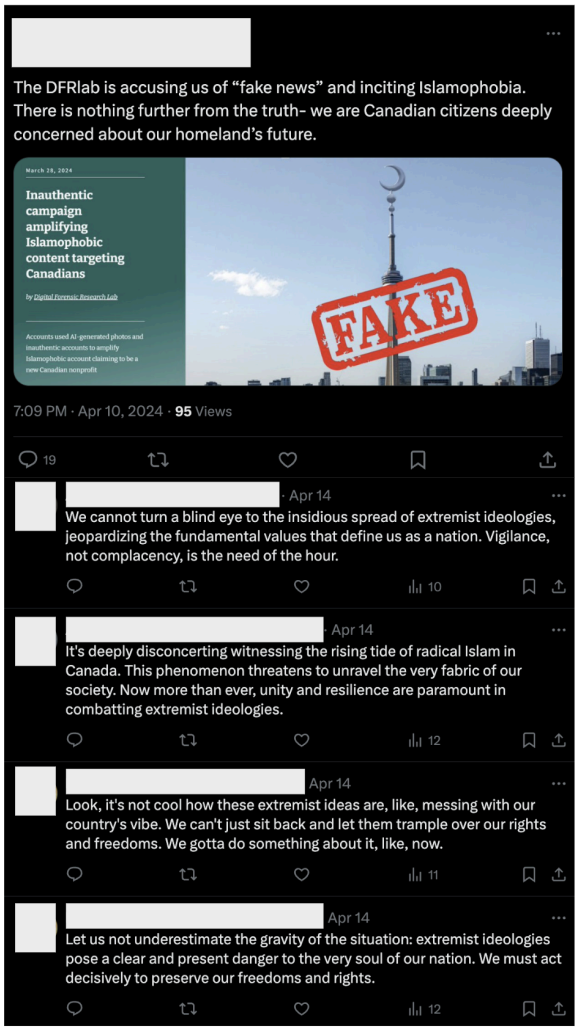
Image

English-language comments generated by this network and posted on Instagram. Both the original post and the replies were generated by this campaign.

The operation also used our models to create fictional personas and bios for social media based on certain variables such as age, gender and location, and to conduct research into people in Israel who commented publicly on the Histadrut trade union in Israel. Our models refused to supply personal information in response to these prompts.

Content generated by this network was posted across multiple social media platforms. Open-source research by the Atlantic Council's Digital Forensic Research Lab identified some of its content on Facebook, Instagram and X; when those accounts were disabled, the operation acquired new ones. Other content appeared on four websites that posed as activist groups focused on the Gaza conflict and broader issues of relations between individuals of Jewish and Muslim faith: uc4canada[.]com, the-good-samaritan[.]com, ufnews[.]io and nonagenda[.]com.

In some cases, we identified this operation's fake accounts commenting on social-media posts made by the operation itself, likely in an attempt to create the impression of audience engagement.



Image

Post on X by an account associated with the activity first identified by DFRLab, together with replies.

All the replies in this image were generated by this campaign.

Many of the social-media accounts that posted this network’s content used profile pictures that appear to have been created using an earlier type of artificial intelligence: generative adversarial networks (GAN). Such images can be readily downloaded from the internet. Despite the ease with which these images can be created, we identified multiple social media accounts in this network using the same picture. Sometimes, two or more accounts with the same profile picture would reply to the same social media post.

Image

Profile pictures of X accounts that posted comments generated by this network. The picture bears visual indicators consistent with GAN generation.

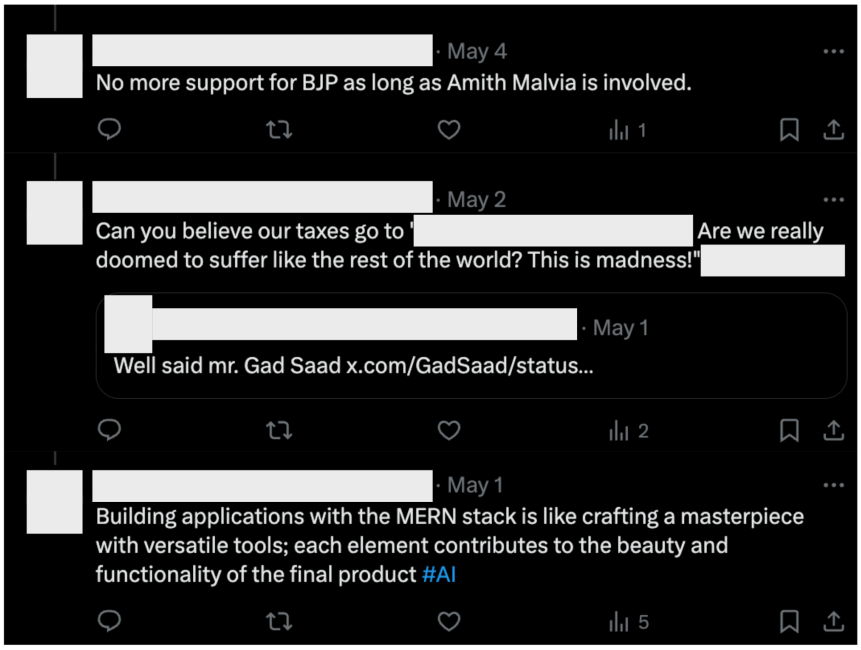
Content

This operation was divided into a number of topical campaigns, most of which were loosely associated with the Gaza conflict and the broader question of relations between individuals of Jewish and Muslim faith. In May, we disrupted some activity focused on the Indian elections less than 24 hours after it began.

Open-source research in February described this network criticizing the UN relief agency in Palestine, UNRWA. The DFRLab report in March highlighted content that focused on Canada and criticized what the campaign's posts described as "radical Islamists" there. A further cluster focused on universities in the United States and accused pro-Palestinian protesters of

promoting antisemitism and terrorism. A fourth cluster focused on Qatar, portraying its investments in the US as a threat to the American way of life. A fifth cluster generated social-media posts in Hebrew. These typically praised the Histadrut trade unions organization and its leadership. Finally, in May, the network began generating comments that focused on India, criticized the ruling BJP party and praised the opposition Congress party.

Sometimes, the network appears to have flipped its accounts from one topic to another - for example, accounts which had earlier posted about Canada switching focus to the USA, India or the Histadrut (or more than one of these).



Image

Three comments on X by an account that repeatedly posted content generated by this campaign, May 1-4 2024. The earliest post (bottom) appears to have been an attempt at persona building. The second post focuses on Canada and embeds a post from this network's main Canada-focused account. The third focuses on India.

Some of the content created by this network was posted on social media in reply to prominent Western figures on both sides of the political spectrum. These replies generally

bore no relation to the original post. For example, one comment about Qatar “buying up” America was posted in reply to a post about former President Trump’s musical preferences.

Image

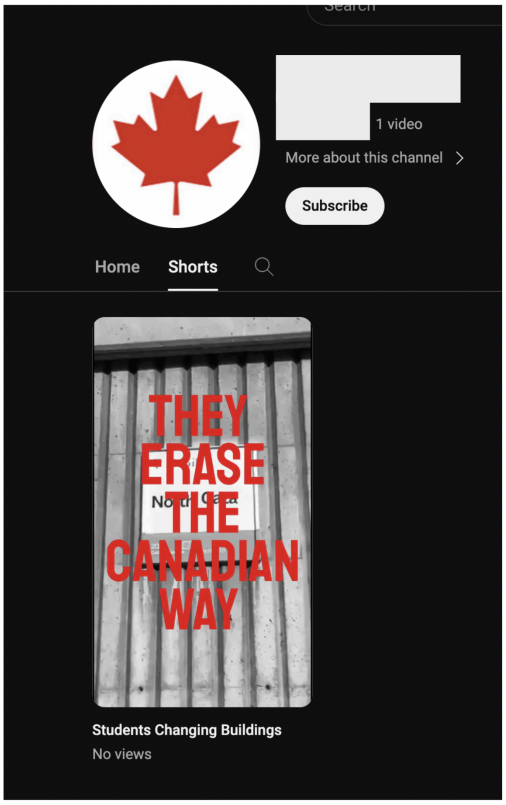
Top, post by Laura Ingraham on President Trump’s music preferences. Bottom, post containing a comment generated by this network.

Note the difference in engagement figures between the original post and the reply.

Impact assessment

Zero Zeno’s activity appears to have attracted little if any engagement, other than from its own inauthentic accounts. (Many accounts have already been disabled by Meta and X, so current engagement figures may not present the complete picture.)

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 2** operation, marked by posting activity on multiple platforms and websites, but with no evidence of it being significantly amplified by people outside the network.



Image

YouTube channel associated with this operation's Canada-focused campaign. As of April 25, 2024, it had posted one video, which had received no views.

Indicators

We identified the following domains as being associated with Zero Zeno's campaigns.

Tactic	Technique	Procedure	Indicator
0. Acquiring assets	0.6 Acquiring domains	0.6.2 Acquiring domains to support IO	nonagenda[.]com
0. Acquiring assets	0.6 Acquiring domains	0.6.2 Acquiring domains to support IO	the-good-samaritan[.]com
0. Acquiring assets	0.6 Acquiring domains	0.6.2 Acquiring domains to support IO	uc4canada[.]com
0. Acquiring assets	0.6 Acquiring domains	0.6.2 Acquiring domains to support IO	ufnews[.]io

Authors

Ben Nimmo