

OpenAI bans Russian and Iranian influence operations run through fake journalists

Source copy kept by wheresthe.ai for the story "OpenAI bans Russian and Iranian influence operations run through fake journalists" (<https://wheresthe.ai/d/openai-false-front-influence-operations-2026-10/>).

Source: <https://openai.com/index/disrupting-ai-enabled-false-front-operations/>

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

SHA-256 of the saved text: 6b10a41c2034b6802ec6a3371420650c5a1b475daf15d177f25e9e188aab5220

Over the past two and a half years, we've reported on many of the ways in which threat actors tried to use our models to conduct cyber attacks, covert influence operations (IO), scams, and other violations of our Usage Policies. We publish these reports to inform regulators, our industry peers, and wider society of how we see threat actors trying to leverage AI, and to shine a light on the vulnerabilities they try to exploit.

We recently banned two IO—one from Russia and one from Iran—that used our models in combination with more traditional techniques and technologies to support sophisticated "false front" entities. They used those entities to launder geopolitical, conflict-related messaging into their target audiences. The Iranian operation included a stable of seven "journalist" personas which it used to pitch long-form articles to small and medium online outlets around the world; the Russian operation appears to have co-opted unwitting people in Latin America to run a "think tank" on the ground.

Both operations were multi-dimensional, trying a range of different tactics to achieve their goals. As well as long-form articles, the Iranian operation generated batches of social media comments, generally on topics related to the US-Iran war. As well as controlling the "think tank," the Russian operation created fake "leaked" documents and audio scripts, some of which we identified being spread online. Both operations also made heavy use of AI to draft internal reports (the Russian operation did this more than anything else); in both cases, the actors used questionable or outright deceitful methodologies to exaggerate the operators' effectiveness.

What is most striking about these operations is that they closely resembled complex influence operations of the pre-AI age, but used AI to make some of the workflows easier. The journalist personas run by the Iranian operation bore a family resemblance to the fake journalist "Alice (opens in a new window) Donovan (opens in a new window)," a front for Russian military intelligence whose articles were published by a range of Western outlets in 2016-17. In 2020, individuals associated with past activity by the Russian Internet Research Agency ran a fake "news" outlet called "PeaceData (opens in a new window)," which co-opted unwitting journalists around the world into writing for it.

These operations were also unusual for their potential reach. Using the IO Breakout Scale (opens in a new window), which rates IO on a scale of 1 (lowest) to 6 (highest), we would assess the Russia-origin operation as belonging in **Category 5**. This is the first Category 5 operation we've disrupted since we began our reporting. The Iran-origin operation reached **Category 4**. Both managed to land their content (not all of which was generated from our models) in mainstream media outlets, rather than simply posting it on social media. This is consistent with a pattern we've observed across the 30 covert influence operations we've exposed in the last two and a half years: the operations which try to land their content in real media outlets, rather than relying on fake social media distribution, tend to have the highest potential reach and impact.

Matrix of the 30 IO we've exposed since early 2024, assessed by primary distribution method (social media, operation-run website, external publications) and score on the Breakout Scale. Operations which used more than one distribution method, such as running a website and promoting it on social media, are classified by the method which appeared to constitute the core of the operation.

AI can give such false-front operations greater scale, efficiency, linguistic fluency, and editorial ability. The operators can use these advantages to exploit unsuspecting victims, such as employees or editors, and plant their content in front of audiences who have no idea who was behind it, or what their motivations were. These operations also demonstrate the complexity and sophistication of threat actors' use of false personas and entities to achieve their goals. Understanding the vulnerabilities that threat actors exploit in these efforts, and how those vulnerabilities can translate into audience and reach, will be essential to maintaining robust defenses across organizations.

But the covert nature of these false-front operations also makes them particularly vulnerable to responsible disclosure. Both “Alice Donovan” and “PeaceData” ceased their activity after they were exposed. That’s why our goal in reporting these false-front operations is to make further research and disruption easier—and make continuing the operations harder.

Russia: Operation “Dark Clark”

Russia-origin actors running influence campaigns across Latin America.

Actor

We banned a cluster of ChatGPT accounts that originated in Russia. They used ChatGPT for a range of tasks associated with covert influence operations targeting countries across Latin America. Much of this activity appeared aimed at undermining Ukraine’s reputation in the region, but some appeared aimed at influencing local political outcomes, especially in Argentina and Bolivia. Most of the operators prompted in Russian; one prompted in Spanish, but nevertheless appeared to be located in Russia. We have shared information on this case with the relevant authorities.

The operators used ChatGPT to perform three main tasks: writing internal reports on their activities (and other people’s activities which they could plausibly take credit for), creating content for their operations, and drafting reports on the performance of a self-described “research platform” in Latin America called the Social Research Center (SRC). They appear to have controlled the SRC via a fake persona named “Mia Clark”; in honor of the name, we have nicknamed this operation “Dark Clark.” Since we do not allow access to our models from Russia, they used VPNs to connect to our services.

In their internal reports, the operators claimed to have spread fake stories across Latin America to undermine Ukraine or local leaders. Some of these fakes have been attributed by open-source(opens in a new window)researchers(opens in a new window) to a Russian entity known as “Politology” or “La Compania,” a reported successor to the Wagner Group and other entities founded by Russian oligarch Yevgeniy Prigozhin. The operators also asked our models to translate or explain public reporting about Politology and Wagner, far more than asking about any other Russia-origin networks.

This operation was unusual in two ways. First, it appears to have successfully co-opted individuals in Latin America to work for the SRC. The Russian operators’ reports on the SRC referred to decisions over issues such as pay scales, hiring, and firing, suggesting that they controlled the entity, rather than cooperating with it. The available evidence indicates that the SRC’s employees in Latin America were not aware that they were working for a Russian group. Unlike another Russia-linked “think tank” that we recently exposed, the SRC appears to have produced a majority of original content via its co-opted staff. This stands out as the most complex attempt to run a front identity that we’ve disrupted over the past two and a half years.

Second, some open-source evidence suggests that Dark Clark’s fakes spread widely enough to provoke fact checks(opens in a new window) and official denials(opens in a new window), indicating a degree of penetration which goes beyond any of the influence operations we disrupted over the past two years. One fake that the operation claimed to have planted in Peru even fed into public tensions between Ukraine and Poland.

Internal reporting

The main way the operators used ChatGPT was to draft and update internal reports to an unknown superior. These regular reports described efforts to conduct covert influence campaigns across Latin America. They encompassed three main workstreams: efforts to denigrate Ukraine and undermine recruitment for the Ukrainian Armed Forces; efforts to interfere in the domestic politics of certain countries, especially Bolivia and Argentina; and management of the SRC. In the majority of cases, we did not observe the threat actors using our models to create content for the campaigns, only to report on them (the few exceptions are described below).

The reports were rich in tactical detail, shining a light on how the operation sought to undermine Ukraine and some national leaders, notably the presidents of Argentina, Bolivia, and Ecuador. Some described efforts to trick locals into carrying out activities that the operators could weaponize; others described fake “leaks” that the operators claimed to have spread.

In two cases, the evidence ties the operators to public reporting on “Politology.” The operators reported that, in 2024, they ran two campaigns targeting Argentina’s President Javier Milei. First, they said they spread a false report that Milei had bought Cartier jeweled collars for his dogs. Second, they claimed that they paid local actors

to post anti-Milei graffiti in Buenos Aires. Based on leaked documents, both these claims have been publicly(attributed to “Politology” and “La Compania.”

Many more fakes have not been previously tied to Russian IO. For example, the operators claimed that in May 2026, they created a fake email address purporting to come from the Regional Directorate of Education in Lima, Peru. They used this to instruct schools in the district to hold events dedicated to Ukraine on the national Day of Cultural and Linguistic Diversity (May 21). The emails included explicit instructions to reference, among others, controversial twentieth-century Ukrainian nationalist Stepan Bandera, who is seen by some Ukrainians as an independence figure, but associated in Russia and Poland with fascism, wartime collaboration, and atrocities. According to the operators, some schools replied to the fake email address, confirming that they had held such events and even providing pictures.

The operators then claimed to have planted stories about the events in the media in both Peru and Poland, alongside allegations that Ukraine was “exporting” ultra-nationalist ideologies, triggering outrage. Open-source searches identified stories that matched this claim in the PeruvianandPolishpress, and an English-language publication in Hungary (some of the articles have since been deleted). Somearticles included a reaction from a Polish Member of the European Parliament proposing that people who showed “anti-Polishness” be declared *persona non grata* in Poland. The Russian fake thus both fed on, and fed into, historical tensions between Ukrainians and Poles.

Similarly, in June, the operators claimed they used a different fake email address to trick schools in Ecuador into holding a ceremony pledging allegiance to President Daniel Noboa and to Erik Prince, former head of private military contractor Blackwater. The operators claimed that the incident provoked outrage in Ecuador and put pressure on the government to deny the fake, thus amplifying it to a nationwide audience. Again, open-source research identified mediacoverage in the Ecuadorianpress that closely resembled this claim, and even a detailed rebuttal by Ecuador’s Minister for Education.

The operators used a range of techniques to underpin their false stories. According to their internal reporting, they spread two different fakes targeting Ecuador in March. One used fake audio attributed to Ukraine’s consul in Ecuador, in which he was alleged to have made disparaging comments about Ecuadorians. Our open-source investigation identified a post on TikTok which matched this precise claim; it appeared to receive only limited engagement. The other described a fake video clip spread on X and TikTok under the logo of a genuine news channel, alleging that Noboa’s government was recruiting young men to fight in Ukraine. Fact checkers in Ecuador described a campaign that matched this description in early April.

TikTok video posted on April 8, 2026, with the faked audio claimed by this operation and attributed to the Ukrainian Consul.

Another internal report described a fake they spread in Bolivia in late May, at the height of anti-government protests. The operators described the goal of this campaign explicitly as to exacerbate the crisis around the protests. The campaign featured a fake audio recording of a worker at the state water company, EPSAS, saying that the government was going to shut off water to the administrative capital, La Paz, and declare a state of emergency. Our open-source research identified a debunk of the matching claim and an official denial from EPSAS.

Other internal reports included fakes that we were unable to corroborate from open sources, possibly because they had already been taken down, or failed to spread. For example, the operators claimed to have spread a story in January alleging that a Brazilian influencer and model had been trafficked into Ukraine by representatives of the Ukrainian Embassy, and had been killed in Kyiv. In June, they claimed to have spread another fake audio clip in Bolivia, this time purporting to come from an employee of the Central Bank of Bolivia, warning that the Bank was about to limit cash withdrawals. They also claimed to have spread a faked letter from the country’s hydrocarbons agency, warning of restrictions on fuel sales to private cars.

In other reports, the operators asked our models to help identify incidents in the target countries that they could take credit for, even if they had not been involved. For example, in Argentina, the operators claimed that the Argentinian Foreign Minister had quoted their operation in a United Nations committee meeting on the Falkland / Malvinas islands. The operators took credit for seeding the arguments that the principle of self-determination does not apply to the islands, and that the UK maintains a disproportionate military presence there. In fact, both

arguments have been part of Argentina's official position for many(opens in a new window)years(opens in a new window). Similarly, in June, Brazilian media reported(opens in a new window) on the case of a Brazilian citizen who had travelled to Ukraine to join the army, been captured by Russian forces, and then made a televised statement claiming that he had been tricked into fighting. The ChatGPT operators tried to claim that this report reproduced their operation's messaging; in fact, it quoted the captive's own video(opens in a new window). This suggests an attempt by the operators to run an influence operation targeting their own employers, rather than any external audience.

The "research platform"

Some of the operators' reports referred to a self-styled "research platform" called the Social Research Center (in Russian, Центр социологических исследований). According to the SRC's website, it focuses on the Indian diaspora in Latin America, and relations between India and the countries of the region.

The internal reports suggested that the operators controlled the think tank, rather than merely working with it. They referenced, for example, hiring and firing decisions, pay scales, staffing plans, and ongoing research projects. Crucially, they also stated that the operators had created a fake persona, named "Mia Clark," to run the operation's social media activity and engage with staff on the ground in Latin America. The available evidence suggests that these on-the-ground employees had no idea that they were working for a Russian operation, and conducted research tasks in good faith. This use of unwitting cutouts strongly resembles the earlier Russian operation also linked to Prigozhin's world, "PeaceData," which Meta exposed(opens in a new window) in 2020.

To judge by the Russian operators' reports, their staff on the ground in Latin America conducted interviews with experts and commentators across the region, and then drafted research papers based on their findings. The topics ranged from a broad analysis of public opinion on the BRICS group, to a detailed comparison of the Brazilian economy under presidents Jair Bolsonaro and Luiz Inácio Lula da Silva. Sometimes, the Russian operators drafted status reports on the research projects before they had been published, indicating close engagement in the production process. We identified well over 60 articles on the SRC website, the great majority of which appeared to be original compositions. This sets Dark Clark apart from the Russia-origin operation we exposed recently, which also ran a think tank as a front, but plagiarized most of its content; it suggests an intention to build a more durable front entity, with contacts throughout Latin America.

To promote the SRC's work, the Russian operators claimed that they had tried to create a network of social media assets. According to their reports, this social media activity ran into problems when operators in different parts of the world tried to access the same accounts, leading to account restrictions. Traces of this can be seen in the open-source record: for example, the transparency settings of the main SRC account on X showed a German location (possibly as a result of VPN use) but a connection via the Russian Federation App Store, while the transparency settings of one of its Instagram accounts showed an admin location in Venezuela. None of the three SRC-branded Facebook accounts that we identified had enough followers to trigger admin location settings.

Left, X account linked to the Social Research Center. Note the transparency setting that shows the account connected via the Russian Federation app store. Right, SRC Instagram account: note the Venezuela location.

The operators reported particular problems on LinkedIn, where they said they created an explicitly fake account to represent SRC, but then struggled to build a network for it. According to their reports, the solution was to create more fake accounts to follow the first one and make it look more convincing. As of August 17, a LinkedIn account with the same Social Research Center branding counted 961 followers and claimed 200-500 staff, but only listed two employees.

Creating fakes

Occasionally, as well as using AI to update reports, the operators did ask our model to help design and create texts that they could use in their operations: for example, asking for help in making sure that their fakes used the correct accent, vocabulary, and institutional style for different countries. This was a relatively small proportion of the overall workload, perhaps because the operation appears to have included at least one native Spanish speaker from Latin America, and therefore had less need of AI language support.

For example, on one occasion, one of the Russian-speaking operators asked the model to help draft, in Russian, a letter which would implicate Ukraine's honorary consul to Panama in corruption. The Spanish-speaking operator then asked the model to translate the letter into Spanish, and to generate an audio script to accompany it. A photo that matched the text of the letter, with a voiceover that matched the operator's script, was posted on TikTok by an

account that purported to be a news outlet, but with the logo of Panama's TVN Noticias(opens in a new window). The TikTok channel only had one follower, and stopped posting after it planted this fake.

TikTok post showing the letter whose text was generated by this operation, with a voiceover script likewise generated by the operation.

On another occasion, the Spanish-speaking operator in Russia gave the model a Russian-language script that claimed Noboa had insulted poor people in Ecuador, and asked the model to translate it into Ecuadorian Spanish. On a third, in March, the same user asked the model to proofread a draft Spanish-language contract between a company called "International Security Solutions FZE" and an individual in Ecuador, purporting to provide infrastructure services in Ukraine. The user's questions focused on how to improve the language and layout, including asking which words should be in bold.

The user's prompts about the contract finished once the wording was finalized. Open-source research, however, shows that a picture of the "contract" was subsequently circulated on social media(opens in a new window). It was used to bolster the claim that Noboa's government had worked with a front company to trick Ecuadorians into being recruited into the Ukrainian army. The fake spread widely enough to trigger a fact check(opens in a new window) in early April: the fact checkers concluded that the company "does not appear in the registry of Ecuador's Superintendency of Companies, Securities, and Insurance".

Impact

The impact of this operation requires particularly careful assessment. As noted above, one of the operators' main approaches was to take credit for activity which had nothing to do with their operation. As such, their own claims of impact in their internal reporting cannot be taken at face value. Similarly, some of the fake stories that they claimed to have planted did not show up in open-source research, potentially because they were spread in non-public channels, already taken down, or not actually posted.

However, open-source evidence allowed us to corroborate some of the claims that they made. Some of the fakes that the operation claimed to have spread were subsequently reported by fact checkers and debunked by governments, suggesting that they had spread widely enough to merit a response. In other cases, we were able to identify content online that closely matched the content generated by this operation. Most notably, the operation's claim to have tricked a Peruvian school into holding an event that portrayed Bandera was subsequently reported on in both Latin America and Europe, and appears to have led to comment from at least one Polish MEP.

Using the IO Breakout Scale(opens in a new window), which rates IO on a scale of 1 (lowest) to 6 (highest), we would assess this operation as belonging in **Category 5**, with evidence that suggests that it led to public comment by politicians in a number of countries. This is the first Category 5 operation we've disrupted since we began our reporting.

Iran-origin operation seeding long-form articles under deceptive bylines in online news outlets.

Actor

We banned a cluster of ChatGPT accounts that originated in Iran. They prompted in Persian, and generated content in Persian and English. The actors obscured their location by using VPNs to access ChatGPT.

The operation used our models to refine long-form articles and editorial pitches, generate social media comments, and refine internal reports. The operators used seven different bylines to pitch their articles, so we have dubbed their campaign "Bogus Bylines." The long-form articles focused on the US-Iran conflict. The majority of the comments likewise dealt with the conflict; a minority dealt with US politics, typically in the wider context of the conflict.

The apparent impact of their main workstreams varied significantly. We identified almost 100 articles published or syndicated under the operation's bylines across roughly a dozen online outlets around the world. These were small to medium outlets, generally focused on international affairs, geopolitics, and events in the Middle East. By contrast, none of the social media comments appears to have generated substantial engagement in terms of likes, shares, or further replies. In fact, the operators' internal reports used a deceptive calculation to measure their social media impact, possibly in an attempt to inflate the operators' apparent effectiveness.

The overall activity appeared consistent with a commercial actor running a for-hire influence campaign, but we are unable to identify the particular actor involved. We have shared information on this case with the relevant

authorities.

Planting articles

This operation's most far-reaching activity consisted of asking ChatGPT to review and refine long-form, English-language articles about the geopolitics of the US-Iran conflict. Typical prompts asked the model to assess an English-language draft against the submission criteria for a specific online outlet, and suggest edits to make sure it complied. Once the operators approved the edits, they asked the model to generate an email that they could use to pitch the article to the editor of the online outlet in question.

These pitching messages were drafted using seven different fake bylines: Ervin B. Hoskins, Noah Lamington, Sophia Gonzalez, Michael Harrison, Ericka Feusier, Jenny Williams, and Alice Johnson. According to the biographies that were posted online alongside some of these articles, each persona purported to be a Western journalist (usually American, though the Noah Lamington persona's bio claimed that they were in New Zealand), often with a focus on international politics, the Middle East, human rights, or conflict. For one of the personas, "Michael Harrison," the operators asked the model to write the bio itself. (Of note, in March Meta disrupted (opens in a new window) an Iranian IO that also featured a "Michael Harrison" journalist persona.)

Some of these personas were backstopped by accounts across a range of social media platforms. The Michael Harrison persona, for example, was backstopped on Medium, X, and Facebook. The Ervin B. Hoskins persona featured accounts on X, Facebook, Instagram, and Threads. This persona's bio claimed to be "an American freelance writer", but social media transparency settings located the accounts in Iran.

Top, Instagram bio for the Ervin Hoskins persona. Bottom left, X account transparency information, showing the use of the West Asia Android app. Bottom right, Instagram account transparency information, showing location in Iran. Meta banned the Instagram account in August.

Using open-source investigation techniques, we identified almost 100 articles that had been published under one or other of the operators' bylines, across over a dozen different publications. These publications generally focused on international affairs, geopolitics, and events in the Middle East, sometimes from an antiwar or progressive point of view. The earliest article we identified was published in July 2025, and the latest in October 2026. The frequency of articles greatly increased after the outbreak of the US-Iran conflict.

Timeline of articles, by author and theme, July 2025—September 2026. Note the density of publications on US-Iran relations from March 2026 onwards.

Sample headlines of articles published under the bylines linked to this operation between February and July 2026 in a range of online outlets.

At least some of the articles were then promoted by the social media accounts of the outlets in question, giving the Iranian operators a potential way to access audiences that they would not otherwise have reached.

Alongside the articles, the threat actors asked ChatGPT to generate batches of comments on geopolitical or political topics in English or Persian. The actors' typical approach was to input the text or a screenshot of a social media post, and ask the model for a set of replies that aligned with their overall focus. These were then posted online, often within a few minutes of each other. The operators' internal reporting referred to both human and automated posting; we have no evidence to suggest that this automation was carried out with the help of our models.

Tweets in English, left, and Persian, right, generated by this operation and posted in July. The English-language comments replied to an English-language tweet warning that the US was preparing a "large scale assault" on Iran. All the comments in this image were posted in a 10-minute span. The Persian-language comments replied to a Persian-language tweet about US strikes on Iran, and were posted within 13 minutes.

Some of this batch commenting activity consisted of generating replies to social media posts about the operation's own news articles. Here, the operators input screenshots of posts that promoted their articles, and then generated comments that praised or endorsed them, likely to make the articles look more popular.

Facebook post by a local US news Page sharing an article generated by the operation, and comments on that post generated by the operation.

The majority of the comment batches focused more generally on the conflict between Iran, the United States, and Israel. The users asked for comments on topics such as missile strikes, ceasefires, and diplomatic maneuvering,

specifying outputs that portrayed the US as the aggressor, Israel as an unreliable ally of the US, and Iran as a defiant but resilient victim. The operators generated a particularly large number of replies to posts by UK-based satellite network Iran International—a critic of the Iranian regime—generating over two dozen batches of hostile Persian-language replies. A minority of the batches generated replies to social media posts about US political issues, typically also in the wider context of the conflict.

Instagram comments generated by this operation in reply to posts about US politics.

We identified a small number of these replies being posted online; many more did not appear in our open-source searches, potentially because they were either not posted, or posted by accounts that had already been taken down. (For example, three of the journalist personas—Ericka Feusier, Alice Johnson, and Jenny Williams—listed X account handles in their published bios: all had been suspended as of August 2026.) Some of the accounts that did post comments online had Western-seeming names, but showed account locations in Iran.

The transparency settings of the Instagram accounts that made the comments illustrated above, showing that they were located in Iran.

Where we did find these replies being posted, they typically made up a minority of total comments on the post in question, suggesting that the operators were not a dominant voice among the authentic responses. Nor did they appear to have attracted substantial engagement: indeed, many of the likes on the operation's replies came from accounts that were themselves linked to the operation. We have shared information on this case with industry partners, who are best placed to provide holistic assessments on engagement.

Internal reporting

The attempt to make their activity look more popular was not limited to public-facing posts. Alongside their content generation, the operators occasionally asked ChatGPT to refine what appeared to be internal reports about their social media campaign. This typically consisted of providing the model with a series of statistics, and asking it to turn them into polished presentations.

This is not the first operation we've disrupted that used AI to write its own(self-review(opens in a new window)). What was most striking about the Iranian operators' efforts to mark their own homework was the method they used to do so. The main "impact" metric they used was the number of views on the posts to which they replied (_not_ the number of views on the actual replies), counted before, and an unspecified amount of time after, the replies were posted. This calculation did not account for any views which may have been the result of other factors, such as the original posts' popularity or other comments. As such, it greatly exaggerated the operation's impact figures, suggesting that the operators either did not know a more effective way of calculating them, or did not choose to use it.

Impact

In fact, based on the evidence that can be gleaned from open-source observation, the operation's two main workstreams appear to have had very different degrees of likely impact.

The social media replies that we were able to identify generally enjoyed low numbers of likes, views, and comments (in the single or double digits). As noted above, the replies typically made up a minority of total comments on the post in question. Using the IO Breakout Scale(opens in a new window), which rates IO on a scale of 1 (lowest) to 6 (highest), we would assess this part of the operation at ****Category 2****, with comments on multiple platforms, but little sign of breakout or ability to dominate the replies.

By contrast, the journalist personas succeeded in getting their articles published by a range of outlets, some with substantial followings: for example, one of the online outlets that published this operation's articles had almost 2 million followers on Facebook, almost 355,000 followers on X, and over 544,000 on Instagram as of August 2026. We would assess this part of the operation as a ****Category 4****, based on its consistent breakout to a number of media outlets.