ISBN 978-9934-619-66-3

VIRTUAL MANIPULATION
2024
ISSUE
1

Virtual Manipulation Brief

Hijacking Reality

The Increased Role of Generative AI in Russian Propaganda

PREPARED AND PUBLISHED BY THE

NATO STRATEGIC COMMUNICATIONS CENTRE OF EXCELLENCE

ISBN: 978-9934-619-66-3

Report prepared by Osavul
Project managers: Dr. Gundars Bergmanis-Korāts, Max Arhippainen
Contributors: Marija Isupova, Bonnie Golbeck
Design: Inga Ropša

About the Virtual Manipulation Brief

The Virtual Manipulation Brief expands the remit of the NATO StratCom COE's Robotrolling series, first published in 2017. It extends the analysis beyond automated messaging about NATO's presence in the Baltics and Poland to include Russia's discourse about the Alliance more broadly, as well as the online campaigns against Ukraine. Launched in October 2022, the Virtual Manipulation Brief offers a concise roundup of the latest insight into the extent, reach, and influence of social media manipulation.

In our inaugural issue of the Virtual Manipulation Brief, we tracked changes in the Kremlin's communication about its War against Ukraine. We analysed the impact of EU sanctions and tracked how Russian propagandists shifted their operations to Telegram. The second issue highlighted how new generative AI tools can also be a boost for defenders. In the third issue we studied how verified propagandist profiles on Twitter (now X) can benefit from that status. We also dived deeper into the AI tools used for information campaigns, and analysed the activity of Russian propaganda in the Hamas-Israeli war.

The current issue of the Virtual Manipulation Brief was prepared in partnership with Osavul – an AI-powered software platform for information environment assessment. Osavul's technology has been actively used to combat hostile influence operations in Ukraine, across the EU and neighboring countries and in the U.S.

The technical capabilities of Osavul provided a comprehensive coverage of threat actors activities, specifically on Telegram, and automated identification of hostile campaigns against NATO and coordinated groups involved in them.

Riga, June 2024

NATO STRATCOM COE
11b Kalnciema Iela
Riga LV1048, Latvia
www.stratcomcoe.org
Facebook: stratcomcoe
Twitter: @stratcomcoe

Executive Summary

More than two years after Russia began its full-scale invasion of Ukraine, coordinated groups[1] on social media continue to pose a significant threat, now compounded by the use of generative AI content, as detailed in this report.

Our research highlights the mixed use of coordinated groups on social media, automated cross-referencing, and AI-generated content. We have identified 17 coordinated groups of accounts for 344 sources. The Virtual Manipulation Brief 2023/1 noted the limited capacity of social media platforms to manage coordinated inauthentic behavior and the increasing focus on AI in information operations[2]. However, as of early 2024, various AI tools are widely used, potentially amplifying the effectiveness of information campaigns and reducing production and distribution costs.

Large Language Models (LLMs)[3] are actively used to create information threats. We have identified automated groups leveraging LLMs to generate noise and scrape websites to repost political news content. These groups extend their activities beyond websites, incorporating social media accounts for broader content distribution. They employ automated dynamic cross-referencing of network accounts, significantly amplifying the reach and impact of their threats without requiring additional funds.

AI-generated comments on Twitter (now X) and Facebook simulate political discussions on sensitive topics. Attempts to prompt LLMs to respond to visual content have seen some success, posing a growing threat as LLMs become more adept at responding to visual media, potentially overwhelming users who may not verify such content.

Between November 2023 and May 2024, automated comments about NATO accounted for approximately 7% on English-language Twitter (X), 15% on Russian-language Twitter (X), and 38% on Russian-language VKontakte (VK). These figures highlight the diverse application of automated accounts across platforms (see Figure 1[1]).

Coordinated groups also exploit digital platforms' attempts to identify inauthentic accounts, using labeled accounts to their advantage as marks of "alternative opinion" in

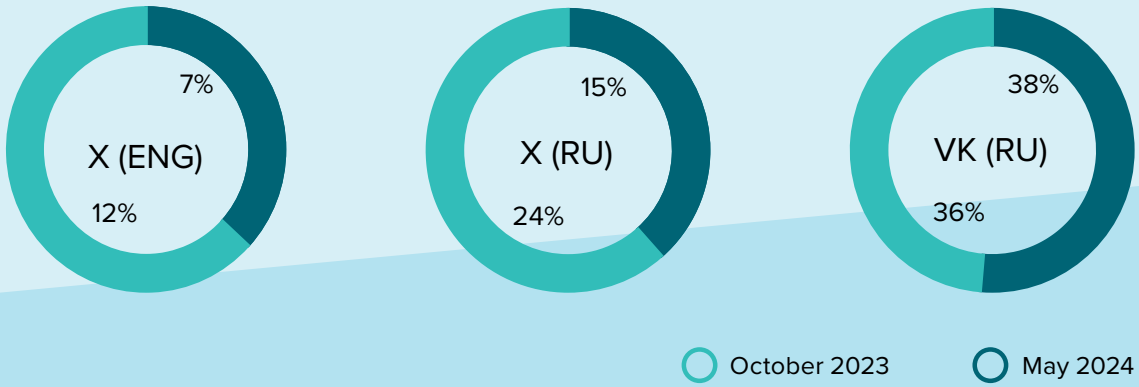

FIGURE 1: *Percentage of automated posts, reposts, and comments about NATO in Nov-May 2024 by platform compared to May-Oct 2023.*

[1] Enhancements to the data collection process, including refinements to the bot detection methodology, may result in minor variations compared to prior reporting periods.

presenting information. This tactic undermines the platforms' strategies against information manipulation.

On March 22, 2024, the terrorist attack at Crocus City Hall, the deadliest in Russia in two decades, resulted in over 10,000 messages across various platforms and websites alleging Ukrainian involvement and Western support in the attack. The terrorist attack at Crocus City Hall had no significant impact on Russia's war in Ukraine and did not noticeably undermine Putin's rule. ■

Hostile messaging about NATO

This analysis delves into Russian and English-language messages referencing NATO from November 2023 to May 9, 2024 on Telegram, Facebook, Twitter (X), TikTok, YouTube, and VK. For research purposes we used a proprietary analysis tool – Osavul[4]. On average, about 200,000 messages[5] in different languages were collected each month, from which key narratives were identified using an AI-based model and grouping messages by common themes.

For Russian-language sources, a spike mentioning NATO was observed on February 27, 2024, when the Lithuanian ambassador to Sweden, Linas Linkevičius, tweeted about Sweden's membership in the Alliance. He stated that 'Russia's previous false accusations that it is surrounded by NATO are now becoming a reality'. Russian-affiliated media distributed this alongside the statement that 'Linkevičius threatened to 'neutralize' Kaliningrad' to at least 6.87 million viewers, a statement that supports Russia's NATO-as-aggressor narrative.

On April 4, 2024, mentions of NATO in both Russian and English increased due to the organization's 75th anniversary. The Russian state and state-affiliated media spread a dossier on NATO compiled by the Russian state news agency TASS. This publication included the NATO leaders' statement from the July 2023 Vilnius Summit, calling Russia 'the most serious and immediate security threat'. Moreover, Russian-linked media actively

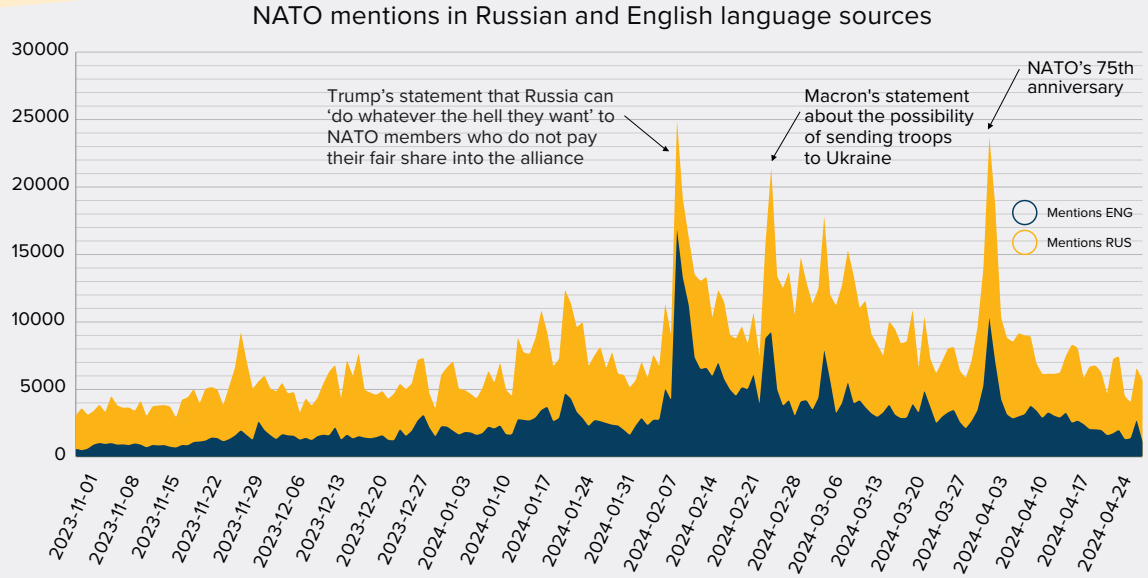

FIGURE 2: Timeline of NATO mentions in Russian and English language sources across Telegram, FB, Twitter (X), and VK from Nov 1, 2023 – April 31, 2024.

disseminated U.S. Secretary of State Antony J. Blinken's assertion that Ukraine will become a NATO member and added comments suggesting that World War III is imminent.

In addition, a further predominant narrative for English-language sources we identified was the statement made by former U.S. President Donald Trump on February 11, 2024. During a campaign rally, Trump stated that he would encourage Russia to 'do whatever the hell they want' to NATO members who do not pay their fair share into the Alliance. This statement was widely propagated by Russian-affiliated media, both state-owned and ostensibly independent with the interpretation that, in the event of war, Europe would be left to fend for itself, devoid of any support from the United States. In total, 669 profiles were involved in the distribution and produced 14.9 million views for the narrative.

Primary narratives against NATO

NATO's Weakness: This metanarrative includes several sub-narratives that portray NATO as ineffective and seek to bolster the image of Russian military superiority.

- **Inferiority of Western Weapons:** Russian and Soviet military hardware is portrayed as superior to Western weapons, with the war in Ukraine cited as evidence. We noted widespread dissemination of reports on the breakdown of a British aircraft carrier intended as the flagship of NATO exercises – 127 messages were spread by 105 actors, and got 4.74 million views. In contrast, Russian media emphasizes the power of Russia's military hardware, such as modernized TU160 fighters, promoting the claim that they are undetectable by NATO troops – this narrative was spread by at least 17 actors in English.
- **Organizational Weaknesses:** Narratives focus on NATO's perceived organizational weaknesses, including empty warehouses, incompetent leadership, and an inability to make quick decisions.
- **NATO as a Non-Threat:** The narrative asserts that NATO poses no real threat to Russia. This was reinforced in late April 2024 by messages from 199 actors about an exhibition of NATO 'trophy equipment' in the center of Moscow – which got 2.51 million views. We observed reports covering this event, alongside retweets of NATO Secretary General Jens Stoltenberg's statements about the exhibition, with commentary on his and NATO's perceived weaknesses. This narrative continued to be disseminated until May 9, 2024, Russia's 'Day of Victory over Nazi invaders'.

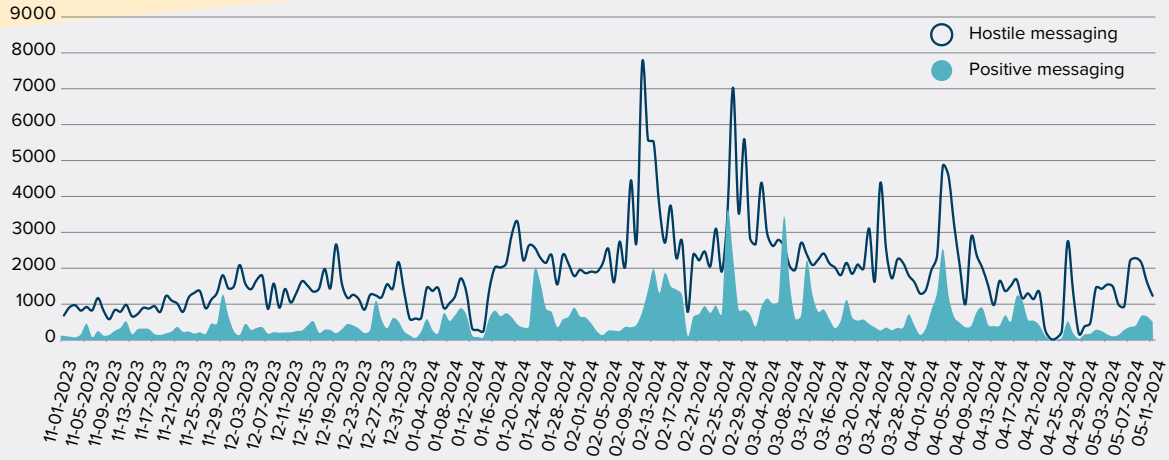

FIGURE 3: Timeline of hostile and positive messaging about NATO across Telegram, FB, Twitter (X), and VK from Nov 1, 2023 – April 31, 2024.

The escalating war against Russia: Pro-Russian actors frequently propagate narratives blaming NATO for the 'instability of the modern world', portraying NATO as the aggressor and Russia as merely defending itself – at least 2644 messages covering this narrative were detected. To disseminate such messages, biased sources often referred to overtly pro-Russian figures; we noted at least 9 such sources, among them: Brazilian journalist Pepe Escobar, American journalist Candace Owens, German businessman and activist Kim Dotcom, Norwegian academic and political scientist Glenn Eric Andre Diesen, and others. In November 2023, a coordinated inauthentic Facebook campaign spread the narrative that NATO continued to expand even after the Warsaw Pact was dissolved – at least 7 different actors spread the same text in the comment section on Facebook. Elon Musk echoed this narrative in March 2024, leading to extensive citations in Russian media. Additionally, in March 2024, pro-Russian sources promoted the narrative of NATO and Ukrainian involvement in the terrorist attack at Moscow's Crocus City Hall. One month later, in April, 2024, Russian Foreign Minister Sergey Lavrov claimed that while Ukraine and Russia were ready to negotiate peace, NATO sought to defeat Russia on the battlefield and was pushing Ukraine into war.

FIGURE 4: An example of a post conveying the "NATO as a colonizer" narrative, illustrating the hijacking of decolonial discourse

Moreover, during our research, we saw considerable activity related to F-35 fighters. Pro-Russian narratives claimed that these jets were part of a dangerous 'network' being established around Russia, including in Ukraine, Finland, the Czech Republic, and Norway – 1097 messages of an increasing threat to Russia from fighter jets produced 5.1 million views.

NATO as a colonizer: Data observed during this research also highlighted the dissemination of several narratives supporting the idea of a NATO-backed genocide in Gaza. These narratives appeared in the comments sections of posts or personal comments on reposted news stories about the military action in Gaza, contributing to the amplification, radicalization, and polarization of the online discourse. Consequently, some groups of reports portrayed Russia as confronting NATO in a struggle for national self-determination. These narratives have hijacked the anti-colonial discourse that has emerged following Russia's full-scale invasion of Ukraine, presenting Russia as resistance against U.S.-backed domination.

The war in Ukraine has already been lost: In the lead-up to the U.S.House of Representatives's vote on military aid for Ukraine inApril 2024, narratives emerged suggesting that the aid would be futile and would not benefit Ukrainians because the war was already lost. It was argued that the aid would be detrimental to U.S. taxpayers, as most of the money would go to military contractors who would profit from the war. Ukrainians were depicted as victims of NATO's desire for war, coerced into continuing the conflict. Concurrently, in April, an exhibition of so-called 'trophy weapons' was held at Moscow's Victory Park, a large memorial complex dedicated to the Soviet Union's triumph over Nazi Germany. This event, presumably, aimed to reinforce the perception among Russians that the war had already been won, despite the new military support allocated to Ukraine in April. ■

AI in Action: The Technical Evolution of Manipulation

Democratizing Propaganda

Coordinated groups on social media platforms remain a significant challenge in countering influence operations. As noted in the Virtual Manipulation Brief from July 2023[6] Twitter (X) has significantly increased opportunities for bot activity due to decisions made by Elon Musk since his takeover of the organization, causing significant alarm. To this end, we identified a case where a user on Twitter (X) echoed a narrative in his tweets that had been launched by a pro-Russian bot network five months earlier. Moreover, the most problematic platform continues to be Telegram, which is not only expanding its influence in the Russian-language space but is also actively working on the creation and development of influence groups aimed specifically at national audiences — American, French, Finnish, and others.

Therefore, it is important to highlight the shift in focus for groups targeting France. In December 2023, Viginum reported that pro-Russian information actors were primarily emphasizing a favorable assessment of the war in Ukraine[7]. However, following President Macron's statements about his readiness to send troops to Ukraine, these groups have shifted their focus to criticizing internal issues within France and targeting President Macron personally. Additionally, we hypothesize that some of the U.S.-based groups we have identified are actively preparing to influence the upcoming U.S. elections.

Our research identified 17 coordinated groups of accounts across 344 sources, primarily on Telegram and Twitter (X). The distribution of these groups is as follows: 67% were Russian-language, 17% were English-speaking, 11% were French-speaking, and 5% were Finnish-language sources. 53% of the groups are active on Telegram, 12% – on Twitter (X), and 35% are mixed – operating son 2 different platforms simultaneously (Telegram and VK, Telegram and TW, or Telegram and WEB).

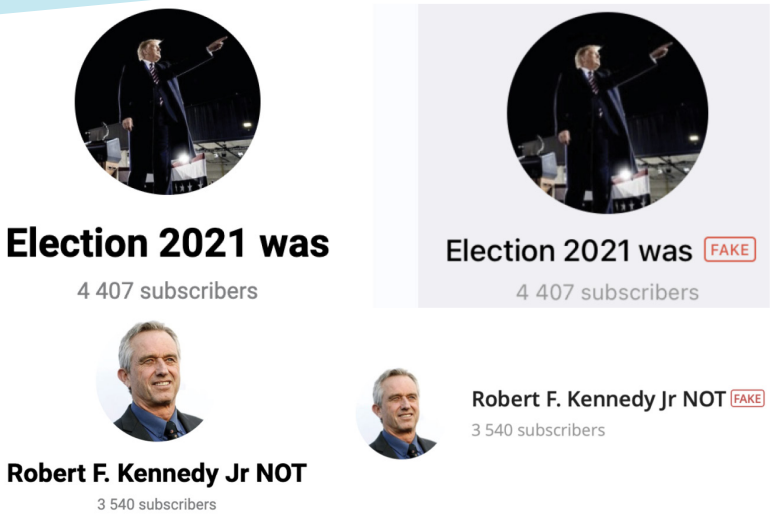

FIGURE 5: Example of how inauthentic Telegram channels impersonating American politicians use creative methods to hide the "Fake" tag

The Finnish and French groups focused on narratives targeting national minorities. The French narratives highlighted the growing national debt attributed to President Macron's military ambitions, and the Finnish narratives suggested that most Finns are unaware of the security risks and consequences of joining NATO, particularly those posed by NATO itself.

One notable English-language group comprises 19 Telegram channels with a total of 293,823 subscribers. These channels impersonate American celebrities and politicians, such as Mel Gibson, Ron DeSantis, and Tucker Carlson. They post right-wing political content to attract a loyal audience and subsequently attempt to scam people financially. The open accounts act as audience magnets, redirecting followers to a closed community called Classified Q Proof (potentially referring to the QANON conspiracy community). During our research, the Classified Q Proof channel was deleted.

Additionally, we identified two similar groups, consisting of 14 and 13 Telegram channels respectively. All 3 groups employ similar tactics: creating fake accounts of famous Americans and creatively using Telegram's fake account tags to make them appear as part of the original channel name. It was used at least for 4 accounts in different groups, 4 more groups were deleted by Telegram during the research because of their inauthenticity. While the first and largest group redirects users to Classified Q Proof, the other two primarily share links to various websites and Telegram channels about U.S. politics. Some accounts from all these groups mimic the accounts of certain U.S. politicians and their relatives. We hypothesize that all 3 groups are related and could be used to influence the upcoming U.S. elections. Their total audience is 438,492.

The most mentioned narratives disseminated by these 3 groups throughout the research period included the American alleged desire to get rid of a failed system – "to destroy the Cabal network dominated by the DARPA/CIA axis and financial tycoons such as the Rockefellers and Rothschilds." They also spread stories about conflicts involving top Ukrainian generals and President Zelenskyy and about Trump's meeting with the Saudi Crown Prince, Mohammad Bin Salman, who is described in publications, as 'the most prominent figure in the Muslim world', while Biden 'is asleep at the wheel.' Our research further substantiates the concept that hostile actors persist in devising innovative strategies to circumvent blocking and labeling mechanisms, thereby effectively weaponizing the very measures designed to counteract information manipulation online. ■

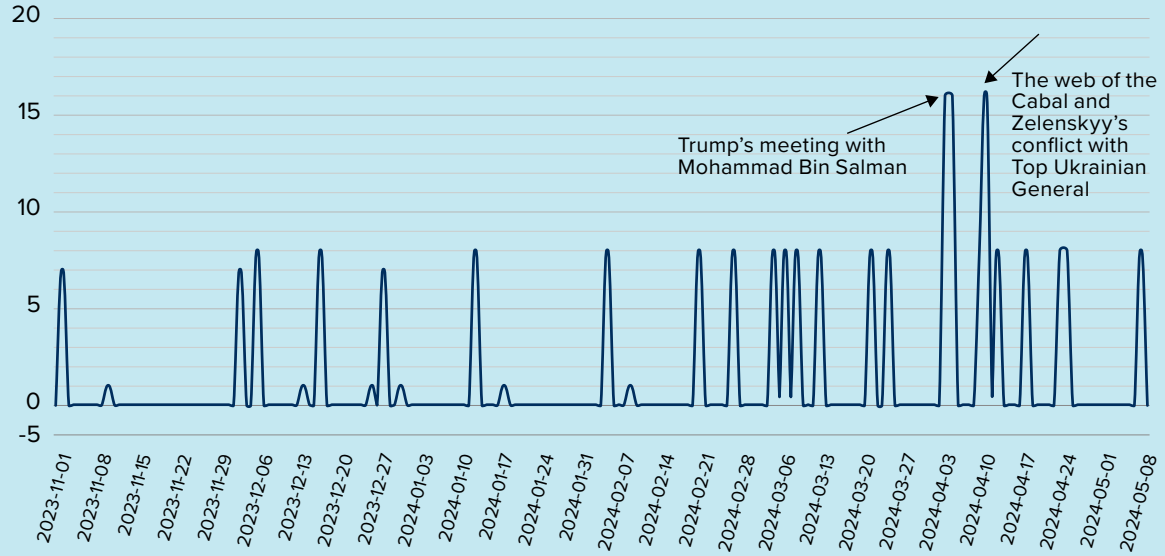

FIGURE 6: The mentions of NATO by three English-language coordinated groups, Nov 1 2023 – May 9, 2024.

How Can We Coordinate Against Foreign Information Manipulation and Interference?

On 23 January, during his keynote speech at the presentation of the 2nd EEAS Report on Foreign Information Manipulation and Interference Threats, Josep Borrell, the High Representative of the European Union for Foreign Affairs and Security Policy, warned that '2024 is a critical year to fight foreign information manipulation and interference' and that 'elections will become the prime target for malign foreign actors[8].'

The concept of Foreign Information Manipulation and Interference (FIMI) is proving increasingly effective. One way to analyze FIMI is through the DISARM Framework[9], a tool that provides a common language for combating disinformation, enabling the defender community to coordinate, share data, analyze, and act in synchrony. DISARM helps facilitate the rapid identification and exchange of data about hostile actor groups, allowing for timely responses to such threats. The effectiveness of this framework is confirmed by successful results of investigations of the Ghostwriter and Doppelganger operations.

Alternatively, there is the RESIST Framework[10], which, while it has a less developed system for classifying and identifying information threats, places a stronger emphasis on timely warnings and responses. Additionally, RESIST offers tools for measuring the effectiveness of these responses.

Both frameworks have significant potential for interoperability, which will allow different government agencies and other stakeholders to streamline threat and knowledge sharing on FIMI. This can be accomplished by standardising the digital exchange of information regarding threats, as well as the methods and effectiveness of responses. Considering the increasing coordination among threat actors, it's crucial to explore this potential now. Such an integration could unify defense capabilities, creating a cohesive system similar to existing comprehensive cybersecurity frameworks. ■

"Considering the increasing coordination among threat actors, it's crucial to explore this potential now. Such an integration could unify defense capabilities, creating a cohesive system similar to existing comprehensive cybersecurity frameworks."

AI-Powered Networks on the Rise

The use of Large Language Models (LLMs) for information threats, as mentioned in the Virtual Manipulation Brief 2023/1[11]*, has already become a reality. Our research uncovered fully automated groups of actors leveraging LLMs to generate arbitrary noise and scrape large websites for reposting political news. These groups encompass both websites and social media accounts that distribute this content. The websites' content is also augmented with AI-generated texts, often resulting in 'AI hallucinations' – nonsense texts with made-up facts. Furthermore, these groups utilize automated, dynamic cross-referencing of accounts within the network, significantly amplifying information threats without requiring additional funding. In addition, AI-generated comments under various publications on Twitter (X) and Facebook have been identified. These comments engage with objections and simulate political discussions on sensitive topics. We also found numerous instances where attempts were made to prompt LLM systems to respond to video and photo content by providing comments based on visual information. Although many of these attempts failed, several successful examples were recorded. This represents a growing threat as LLM systems become more adept at responding to visual content, which is time-consuming for the average user to verify.*

We asked Llava-v1.6-34b[12] *to describe the same photo. The result and details are close; some of the word forms chosen by LLM are identical.*

FIGURE 7: An example of a bot successfully recognizing and responding to images using AI

Among the coordinated groups identified in our study, one notable group comprises 130 Telegram channels and 13 websites, all

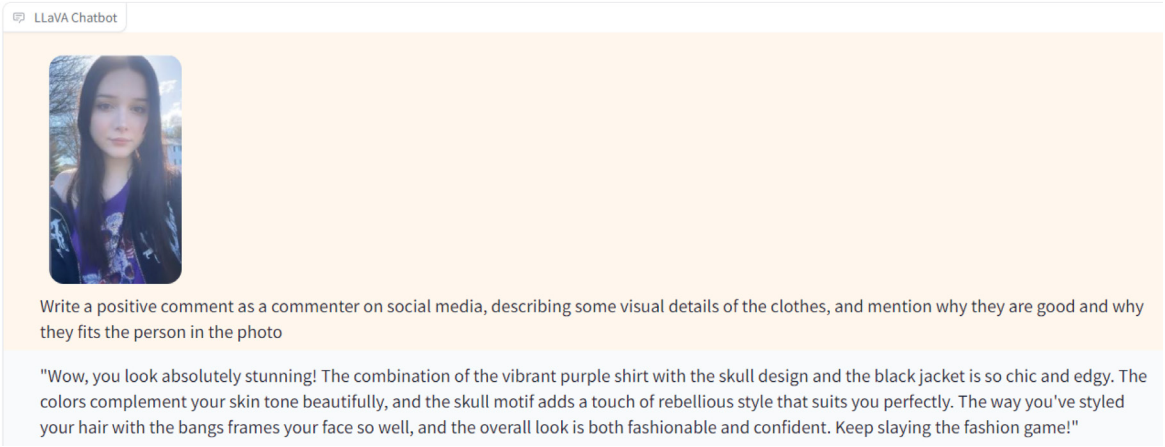

FIGURE 8: LLaVA describes the same photo

in Russian. This group is fully automated, relying on LLMs to generate arbitrary noise and scrape large websites to repost political news. Most of these 13 websites were involved in influence operations in Ukraine[13]*. Additionally, some of these websites are linked to a broader network of web resources that mimic well-known Ukrainian information platforms through their domain and advertisement tag histories*[14].

The cost of creating and maintaining this network is minimal, while the total traffic amounts to approximately 50,000 visitors per month, with 4,675 subscribers on Telegram. Traffic analysis across all sites showed that different domains experienced monthly traffic growth of up to 350%. However, the attendance dynamics were unstable, with sharp increases and decreases, indicating inorganic and thus inauthentic traffic patterns. It is known that at least two domains (pilotekh.com and sitinvest.ne) had their traffic analyzed by the Russian company Brand Analytics[15,16], owned by Palitrumlab LLC[17]. This company has secured over 126 government contracts in Russia[18] and is listed as a state support recipient and contractor for the General Radio Frequency Center (GRFC)[19], a subsidiary of Roskomnadzor[20]. ■

Automated cross-reference

The group utilizes automated, dynamic cross-referencing of accounts within the network. Each Telegram post includes links to two primary sources, which are other Telegram channels within the network. These links are programmatically generated, with pairs rarely repeated, to achieve deep linking. Additionally, the group selectively posts content from major pro-Russian media outlets without proper citation.

Among the NATO references, the group disseminated factual news alongside blatantly anti-NATO narratives promoted by pro-Russian actors. For example, they circulated a statement by former U.S. intelligence officer Scott Ritter, who claimed that 'there is not a single unit of NATO forces that would hold out against the Russians for more than a week', as well as a statement by Mikhail Ulyanov, Russia's permanent representative to international organizations in Vienna, suggesting that 'Finland could be the first to suffer in the event of a hypothetical armed conflict between Russia and NATO.' ■

AI-generated content on websites and social media

Each Telegram post from this group includes a link to one of the 13 websites, all of which share a similar URL structure[21], indicating centralized administration. The group uses AI-generated content for these websites, producing fully fabricated articles based on a common template.

In addition, outside this group, we discovered numerous bots on Twitter (X) posting AI-generated comments under various publications. These bots often revealed their artificial nature through messages where the LLM cited ethical or other reasons for being unable to fulfill requests. By examining these failures, we identified a group of accounts and analyzed their other messages to find successful uses of LLMs and understand the purpose of this activity.

These bot accounts aim to mimic human behavior, reacting to messages with contextually relevant responses, often citing external facts. They frequently comment on issues related to U.S. domestic politics, religion, and economic problems, suggesting a preparatory phase for potential use during the U.S. presidential elections. Moreover, there were

numerous attempts to prompt LLM systems to respond to video and photo content by providing comments based on visual information. Although many attempts failed, several successful examples were recorded.

Generally, AI-generated activity on social media platforms can be divided into two types:

- **Twitter (X) and Telegram:** Accounts support common activities by reposting news or popular posts, sometimes with additional comments. Such additional comments assessing the situation can polarize and radicalize the discussion, however, they also can make messages without any commentary, to help spread the narrative.
- **Facebook and Twitter (X):** Bots actively engage in discussions, promoting necessary narratives under the guise of personal opinions. These narratives may be perceived less critically and not recognized as part of information operations. Bots also play out dialogues among themselves, presenting different versions of the same narrative or addressing objections. This behavior is often detected on professional media outlet accounts and official profiles. ■

Crocus City Hall: A Case Study in Russian Propaganda

On March 22, 2024, Moscow's Crocus City Hall experienced one of Russia's deadliest terrorist attacks in two decades, revealing significant weaknesses in the Russian security apparatus, already strained by over two years of war in Ukraine.

U.S. officials, speaking anonymously, stated that the National Security Council had warned Moscow about the planned attack[22]. The Kremlin did not address the specific warning, and Russian intelligence asserted that the information was too vague to act on. However, video evidence and reports indicated a delayed and inadequate response by Russian security forces during the attack.

In the immediate aftermath, Russian propaganda shifted the blame to Ukraine and NATO member countries, painting them as key beneficiaries of the terrorist attack. On Telegram, Dmitry Medvedev asserted that ISIS had no real involvement and accused Ukraine and Western politicians, particularly French President Emmanuel Macron, of orchestrating the attack.

Medvedev further suggested that ISIS merely took credit for the publicity while the true organizers remained hidden. He claimed that evidence found on the suspects' phones pointed to Ukrainian officials and alleged that Western leaders, including Macron, supported terrorists. President Putin later reinforced this narrative, stating that the investigative authorities' evidence implicated Ukraine as the main culprit.

Our analysis identified approximately 10,000 messages related to the attack distributed on Telegram, Facebook, Twitter (X), VK, and various websites. The key actors involved were pro-Kremlin media and Russian officials. These messages were amplified by controlled profiles and bot accounts in English, Russian, and German.

By consolidating a unified, deceitful communication strategy, Russia managed to control the narrative and mitigate internal damage from the attack. Contrary to expert predictions, the Crocus City Hall attack did not lead to significant consequences for Russia's war in Ukraine or visibly weaken Putin's rule. ■

Telegram's "Disinfonomics"

On March 31, 2024, Telegram introduced a system that allows channel owners to receive 50% of the revenue from advertising placed on their channels through the Telegram Ads system. Given the substantial online traffic generated by disinformation channels, particularly Russian ones, it is worth examining how much these channels could potentially earn. Telegram does not disclose specific data on how much each channel earns from advertising, how views are converted into ad impressions, or provide statistics on regional or thematic specifics. However, we can make some estimates based on Pavel Durov's statement that 10% of views are monetized through Telegram Ads[23].

To estimate the potential earnings of propaganda Telegram channels during the study period, we collected all messages mentioning NATO from channels with more than 1000 subscribers – a requirement for displaying ads on the platform. We focused on channels marked as compromised in Osavul proprietary analysis tool, indicating their involvement in information operations, disinformation, and

being sanctioned, hacked, or stolen. This resulted in 264,697 messages, generating a total of 4.13 billion views.

Taking the 10% monetization rate mentioned by Durov, we arrive at 413 million views participating in the Telegram Ads program. According to Telegram, the minimum cost per 1000 impressions (CPM) on Telegram is 0.1 Toncoin[24,25], equating to 0.05 Toncoin shared with content creators. Since Telegram's announcement in March 2024, the average cost of Toncoin was €5.59 ($6.08). Therefore, 1000 impressions would generate for the channel owner approximately €0.28 ($0.3). It means that NATO-related content in this research could generate its authors an impressive €115,888 ($125,552).

It is interesting to compare possible ad revenue on Telegram to the one from the Twitter (X) Premium sharing program, mentioned in Virtual Manipulation Brief 2023/2. As estimated by Calculate Buddy's formula[26], 4.13 billion views on Twitter (X) would make the content creator approximately €32,402 ($35,105). Based on this calculation, propaganda on Telegram could be up to 3,5 times more profitable, if other variables are unchanged.

Such a difference in value makes Telegram a highly attractive platform for both new and existing content creators, which in turn can be an impetus not only for the development of the platform itself but also for expanding the activity of hostile actors. ■

Methodology

For this Brief using Osavu analysis tool we identified incidents that exhibited characteristics of information threats, including messages promoting negative sentiment, containing fake or manipulative information, or from previously identified sources of information influence campaigns. Following this, identified incidents were supplemented using the "search by similarity" function integrating similar messages in different languages and platforms. Cross-analysis was then conducted to categorize incidents into thematic blocks highlighting key events, use of identical tactics, and revealing ongoing information campaigns throughout the analysed period. Subsequent steps described the narratives and tactics used to create negative publicity around NATO, with particular attention to incidents that aimed to damage NATO's reputation, without taking into account messages that only mentioned NATO occasionally.

The identification of bots relies on detecting inauthentic behavior as a key factor. The process involves collecting a vast amount of comments from digital platforms. These comments are then scrutinized to identify suspicious groups of similar messages, which may originate from different accounts but share common characteristics. Further analysis is conducted on the accounts behind these messages to determine if they exhibit abnormal posting patterns, share numerous suspicious comments, or act as a part of a network.

We also paid a lot of attention to identifying coordinated groups on various social media platforms and related networks of websites. To identify such groups we analyze the content and behavior of all actors in our database searching for similar patterns in the context of a certain topic (NATO). Once the candidate coordinated group is identified, we calculate metrics and proofs of coordination which help analyze and validate the results. ■

Endnotes

1 Groups of actors that distribute messages in a coordinated manner

2 *Virtual Manipulation Brief 2023/1: Generative AI and its Implications for Social Media Analysis*. NATO NATO Strategic Communications Centre of Excellence 5.6.2023

3 Language models that are notable for the ability to achieve general-purpose language understanding and generation.

4 *Osavul* – an AI-powered security against information threats.

5 Average distribution per platform: Facebook – 19.94%, Telegram – 36.83%, TikTok – 0.06%, Twitter(X) – 39.43%, VK – 3.5%, Youtube – 0.24%. Total amount of messages – 1,911,556

6 *Virtual Manipulation Brief 2023/1: Generative AI and its Implications for Social Media Analysis*. NATO NATO Strategic Communications Centre of Excellence 5.6.2023.

7 *PORTAL COMBAT: A structured and coordinated pro-Russian propaganda network*. The General Secretariat for Defence and National Security (SGDSN), February 2024.

8 *Disinformation and Foreign Interference: Speech by High Representative/Vice-President Josep Borrell at the EEAS Conference*. European Union External Action, 23.1.2024.

9 *DISARM Framework*.

10 Pamment, James (2021): *RESIST 2 Counter Disinformation Toolkit*. UK Government Communications Service.

11 *Virtual Manipulation Brief 2023/1: Generative AI and its Implications for Social Media Analysis*. NATO NATO Strategic Communications Centre of Excellence 5.6.2023.

12 *LLaVA:* Large Language and Vision Assistant.

13 *Internet Archive Wayback Machine.* and *Украинский Бизнес День*

14*BuiltWith®* Pty Ltd

15 According to Similarweb report

16 *Brand Analytics*.

17 *Palitrumlab LLC*.

18 Saby: Палитрумлаб, ООО: *ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "ПАЛИТРУМЛАБ"*

19 *General Radio Frequency Center* (GRFC).

20 Litavrin, Maxim & Frenkel, David (2023): *Inside Russia's internet monitoring. How the censorship agency tracks online activity with the help of tech companies*. Mediazona, 8.3.2023.

21 Websites linked to in Telegram posts:
http://1ubd.com/component/k2/item/69591

http://akcentu.com/component/k2/item/68963

http://balansst.com/component/k2/item/78646

http://biznes-tema.com/component/k2/item/74760

http://biznesfakty.com/component/k2/item/74781

http://blogs-exposed.com/component/k2/item/75607

http://businessinblogs.com/component/k2/item/79821

http://face-n.info/component/k2/item/89468

http://pilotekh.com/component/k2/item/83561

http://sitinvest.net/component/k2/item/70824

http://thepoliticalemporium.com/component/k2/item/89859

http://viptrade.biz/component/k2/item/64573

http://x-informer.com/component/k2/item/83918

22 Harris, Shane (2024): *U.S. told Russia that Crocus City Hall was possible target of attack*. The Washington Post, 2.4.2024

23 *Telegram post*.

24 Toncoin is the principal cryptocurrency of The Open Network (TON) blockchain developed by Telegram

25 *Telegram Ad Platform Explained*.

26 *Twitter Money Calculator – How Much Can You Earn via Twitter Ad Sharing Revenue? (Easy)*. Calculate Buddy 7.1.2024.

Prepared by Osavul and published by
NATO STRATEGIC COMMUNICATIONS
CENTRE OF EXCELLENCE

The NATO Strategic Communications Centre of Excellence (NATO StratCom COE) is a NATO accredited multi-national organisation that conducts research, publishes studies, and provides strategic communications training for government and military personnel. Our mission is to make a positive contribution to Alliance's understanding of strategic communications and to facilitate accurate, appropriate, and timely communication among its members as objectives and roles emerge and evolve in the rapidly changing information environment.

www.stratcomcoe.org | @stratcomcoe | info@stratcomcoe.org