ISBN 978-9934-619-57-1

VIRTUAL MANIPULATION
2023 ISSUE
2

Virtual Manipulation Brief

VERIFIED PROPAGANDISTS AND THE HAMAS-ISRAEL WAR

PREPARED AND PUBLISHED BY THE

NATO STRATEGIC COMMUNICATIONS CENTRE OF EXCELLENCE

ISBN: 978-9934-619-57-1

Author: Rolf Fredheim
Contributor: Markolo Research
Design: Inga Ropša

About the Virtual Manipulation Brief

The Virtual Manipulation Brief expands the remit of 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.

The Virtual Manipulation brand includes NATO StratCom COE's annual social media experiment, which tests the capacity of social media platforms to detect and remove commercial social media manipulation. The platforms' ongoing inability to prevent or remove even a fraction of commercial manipulation does not bode well for their capacity to safeguard their systems against subversion by state-sponsored campaigns executed by devoted teams.

Riga, December 2023

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

This publication does not represent the opinions or policies of NATO or NATO StratCom COE.

Executive Summary

In this issue of the Virtual Manipulation Brief, we identified 117 pro-Kremlin accounts—including notorious Z bloggers—who purchased verified status on X. This meant they could monetise fake news through X's ad revenue-sharing feature. It also boosted their visibility – the newly verified propaganda accounts received more than twice as many views per post, on average.

In October 2023, Kremlin propaganda shifted sharply in response to the Hamas-Israel conflict. Pro-Kremlin accounts played a significant role in propagating disinformation, weaving narratives that linked the conflict to Ukraine. The Russian operation adopted a strongly anti-Israeli stance. It advanced a narrative that the US-backed Israel is responsible for killing babies in Gaza. It was transparently aimed at rehabilitating Russia's image on the global stage and diverting attention from Russia's activities in Ukraine.

In the six months spanning May to October 2023, the conversation surrounding NATO in the Baltics and Poland centred on the NATO summit in Vilnius, the reaction to Wagner troops deploying near the Belarussian border, and a drone attack in Pskov.

We examine VKontakte's evolution into a multi-purpose platform, the Kremlin's 'everything app'. VK has expanded to incorporate features mimicking major Western platforms. It aligns closely with governmental interests and has become an essential tool in Russia's digital infrastructure.

Autumn 2023 brought significant new potential for AI-driven social media analysis. New multi-modal AI can handle complex audio, image, and text input combinations. This makes it possible to do better in-depth meme analysis and analyse videos. In 2024, we intend to apply these tools to analysing content on TikTok and YouTube.

We conclude with an overview of the latest developments in social media manipulation, from the deluge of fake photos claiming to show scenes of suffering in Gaza to the emerging trend of artificial audio content targeting political opponents.

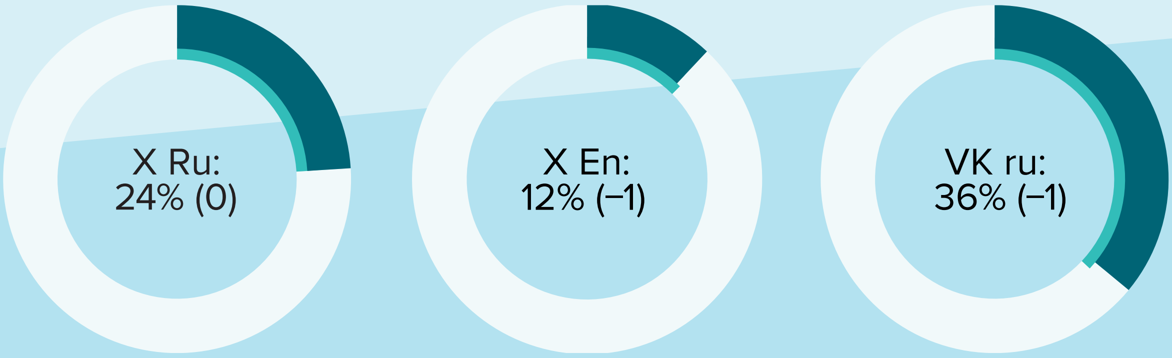


FIGURE 1: *Percentage of automated posts about NATO in May - October 2023 by platform compared to Sep 2022- Apr 2023.*

The volume of conversations about NATO in the Baltics and Poland from 1 May to 31 October 2023, remained consistent with the previous six months. Despite the decision to cut access to the free X API, we observed no significant change in the presence of automated accounts on the platform. In Russian, 24 per cent of posts were from automated accounts, while for English-language posts, there was a minor decline from 12.5 to 12 per cent. On VKontakte, our analysis shows a slight decrease in bot-driven posts, dropping by one percentage point to 36 per cent. Bot accounts are identified as online entities whose activities suggest automation by computer programs or scripts.

On English-language X, we noted a nearly 10 per cent rise in post volumes and a reduction in anonymously operated accounts. Conversely, VKontakte experienced a 15 per cent drop in posts about NATO, with the number of unique users nearly halving. This trend was driven primarily by a reduction in the number of groups and accounts classified as real-name personas. Due to this reduction in authentic activity, the proportion of bot activity increased from 11 to 14 per cent.

Hostile messaging about NATO

For this report, we employed AI to discern trends in pro-Kremlin messaging about NATO in the Baltics and Poland, corresponding to the spikes illustrated in Figure 2. This methodology offers several benefits, including time efficiency, the ability to process many data points, and a more impartial analysis.

The method involves summarising the top-performing posts of each day and then distilling these summaries into a daily digest by language and platform. Finally, these digests are synthesised to provide an overview of the entire period. The insights overleaf are derived from the AI's outputs.

Discussions in English on X typically engaged with NATO's strategies and member state commitments. Russian narratives across X and VK were more critical, often portraying NATO as opposing Russia and questioning the alliance's motives and actions.

The Vilnius Summit on 11 July catalysed a surge in activity on both platforms. English commentary focused on the summit's outcomes and US pledges, whereas Russian discussion amplified strategic Russian manoeuvres and criticism of NATO's enlargement. Earlier in June, comments from former Secretary General Anders Fogh Rasmussen about Ukraine's assurances also sparked commentary on VK.

On 10 August, English-language discussions about NATO on X centred on Poland's deployment of 10,000 troops to its eastern boundary with Belarus as a countermeasure to the Russian Wagner group's presence. Conversely, the Russian Defence Minister accused NATO of concentrating 360,000 soldiers in Eastern Europe, allegedly to invade Western Ukraine.

On August 30, VK users reacted to a drone attack on Pskov airport. Commentators highlighted Pskov's closer proximity to Latvia and Estonia than to Ukraine. Kremlin propagandist Vladimir Solovyov provocatively suggested that the Baltics should be 'f***ing destroyed' following the drone incident at Pskov's airport.

On 8 and 23 October, Russian language X and VK reflected on damages to the cable and pipeline connecting Estonia and Finland. Speculation about NATO potentially limiting Russian access to the Baltic Sea was met with derision and disapproval. NATO was cast as a violator of international laws.

Comparisons:

English content on X focuses more on NATO's obligations and actions, particularly the US's involvement, while Russian content on both X and VK emphasises Russia's strategic stance and criticises NATO. English discussions tend to scrutinise NATO's preparedness and decisions, whereas Russian narratives assert aggression and violations of norms.

English X posts react more to specific statements or incidents, such as the Russian Defense Minister's claims or Zelensky's actions, while Russian content often ties events to broader geopolitical narratives.

Comparing Russian messaging on VKontakte and X, VK discussions delve deeper into speculative and provocative themes. There's a stronger focus on criticising NATO's intentions and actions, often linking them to historical or ideological narratives. VK posts also tend to include more direct and severe criticism of Western and Ukrainian leaders. This difference is due to the comparative absence of Ukrainian voices on VK and the longer messages permitted by the platform.

Main anti-NATO narratives:

Undermining Trust in NATO and its Allies: These narratives typically depict NATO as an aggressor or self-serving in its involvement with Ukraine. In October 2023, propagandists seamlessly extended this narrative to characterise Western support for Israel in its war with Hamas.

Promoting Russian Strength and Independence: Messages underscoring Russian resilience and military success in the face of sanctions are designed to foster national pride and a sense of independence.

Discrediting Ukrainian Efforts and Morale: By stating that Ukrainian forces are demoralised or acting on orders from others, these messages attempt to demotivate opposition and create the perception of inevitable defeat.

Rationalising Aggressive Actions: These messages frame the conflict as a reaction to NATO expansion and Ukrainian hostility, thereby portraying Russian military actions as defensive. This narrative was further highlighted by contrasting the portrayal of civilian distress and alleged transgressions by Israel in Gaza with the sanitised representation of Russian activities in Ukraine. ■

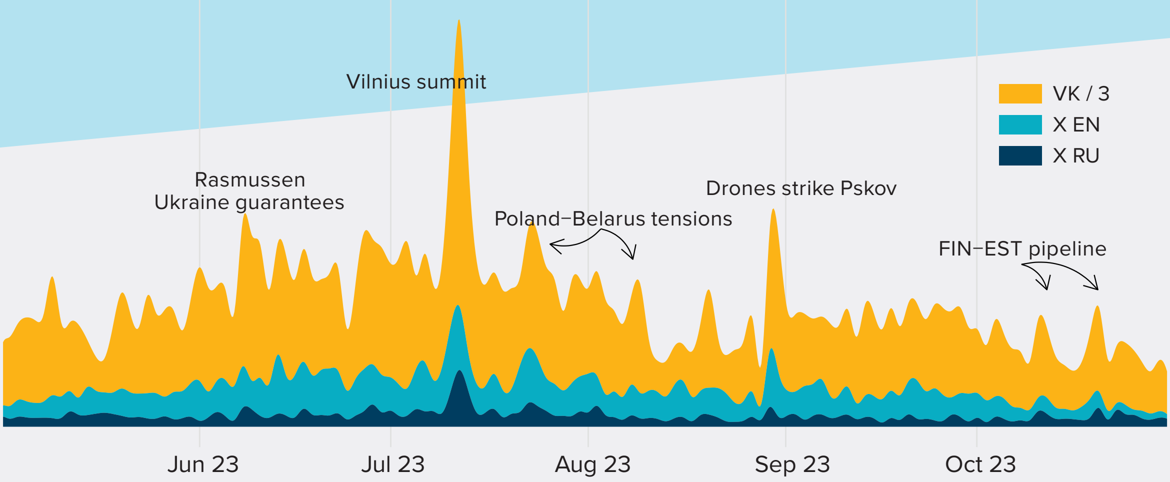


FIGURE 2: *Timeline of commentary on NATO in the Baltics and Poland. VK scaled to ⅓ due to much higher volume.*

X Premium: A Verified Bonus for Propagandists

Our research shows that the policy allowing users to pay for greater visibility on the X has dramatically expanded the reach of pro-Kremlin influencers.

Originally, the verification system aimed to authenticate public figures' accounts, such as celebrities, politicians, and influencers. The verification system helped users distinguish authentic accounts from impersonators or fake ones. In November 2022, X replaced this system with a paid subscription model. In April 2023, Elon Musk announced that the platform's algorithm would prioritise verified accounts.

The new verification system meant professional manipulators could widen their reach. How would the Kremlin's cheerleaders adapt? Would the platform introduce counter-measures to restrict the spread of propaganda? To answer these questions, we identified users who frequently posted about NATO, consistently adopting a pro-Kremlin and anti-NATO stance, and monitored when they obtained verified status. We fine-tuned GPT 3.5, a large language model, to act as our classifier. For each user, we collected a sample of posts and looked for a trend where the classifier thought the output was clearly pro or anti. We compiled a list of the most active accounts, manually verifying their pro-Kremlin positions through recent posts. Using Meltwater, we tracked the acquisition of verification status. Our analysis focused on newly verified accounts under Elon Musk's rules, excluding official government and media accounts with legacy verification. This left 117 newly verified accounts that had paid the $8 for a verified badge.

As Figure 3 shows, the results were dramatic. We observed a 72 per cent increase in retweets and a 109 per cent rise in views on average after verification. The newly verified users included notorious pro-war milbloggers Ayden, Geroman, Rybar, and SpetsnaZ 007. RU. Together, they produced thousands of posts in the period. While most pro-Kremlin accounts are Russian-speaking, most of these verified accounts primarily operate in English. Our sample included 33,500 English tweets about NATO from these verified accounts, compared to only 1,000 in Russian. These stats confirm that X is perceived as more significant for influencing foreign audiences.

Social media platforms reward engaging, outrageous, and polarising content. In the case of X, the platform literally rewards it. On 31 July, Elon Musk tweeted that the platform would share ad revenue with Premium subscribers.

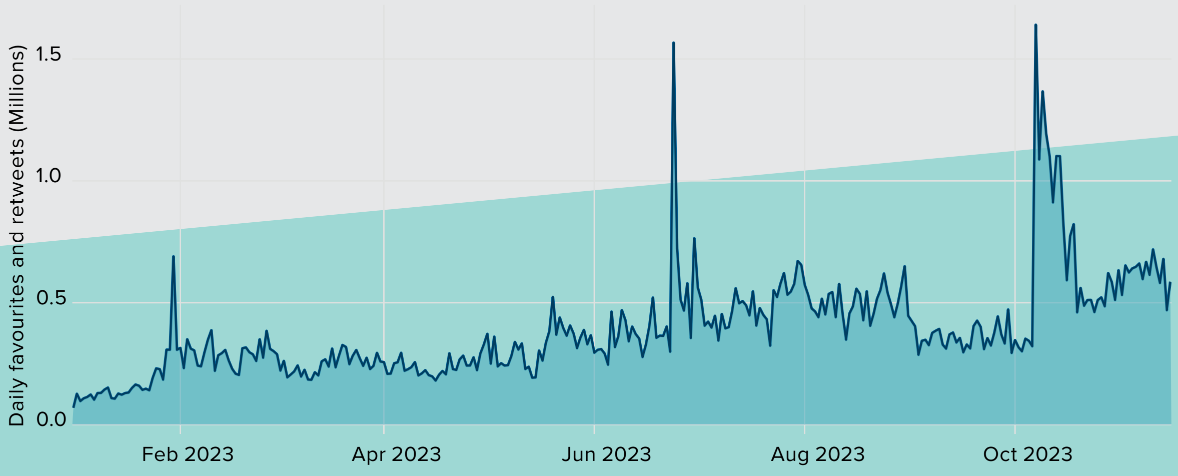


FIGURE 3: *Daily retweets and favourites for newly verified pro-Kremlin accounts on X*

The exact formula for determining revenue is opaque but involves the number of ad views from other subscribers. One estimate places the average return paid to creators per 1M views at $8.5.[1] Altogether, the 117 newly verified accounts achieved an eye-watering 15 Billion impressions, according to Meltwater. Consequently, this sample of 117 pro-Kremlin propaganda accounts could have generated approximately $127,500 in ad revenue. ■

Kremlin propaganda and the Hamas-Israel war

The terrorist attack by Hamas against Israel and Israel's subsequent military incursion into the Gaza Strip sparked international outrage, directed at apparently rising anti-semitism on the one hand and, on the other, the alleged excesses of the military response. There was also a significant spike in fabricated content on social media. Pro-Kremlin accounts played a significant role in spreading disinformation about the Hamas-Israeli conflict.

On Telegram, Dmitri Medvedev posted a photoshopped image of President Zelensky in a beard and a turban with the caption that ISIS 'was a traditional partner of the Ukrainian regime'. Within days, though, the Kremlin's position consolidated around an anti-Israeli stance. A brief survey of RT Head Margarita Simonyan's X posts typifies the official line. Simonyan ignored the terrorist attacks but instead shared conspiracy theories that Hamas used US weapons intended for Ukraine. On 18 October, she falsely accused Israel of bombing a hospital in Gaza and in November of planning 'another genocide'.

Evidence published by *Le Monde* and *Haaretz* appeared to expose a Russian influence operation fomenting anti-semitism in France. In late November, Hamas announced the release of a Russian hostage in appreciation of what it termed 'Russia's support for the Palestinian cause'.

We surfaced the newly verified accounts in the previous section due to their consistent pro-Russian line in the war against Ukraine and messaging about NATO. However, as Figure 4 shows, they sprang to life in early October at the time of the Hamas terrorist attacks and the subsequent Israeli military invasion of Gaza.

The posting frequency surged at the beginning of October, rising to over 6,000 per day compared to an average of 3,500 in September. Since 1 October, mentions of Hamas by these accounts increased sixfold, and the Israeli military threefold, compared to their references to the Ukrainian military. The location 'Israel' was almost three times as frequently mentioned as 'Ukraine' and nearly four times as often as 'Russia'.

Pro-Kremlin accounts immediately linked Hamas' actions to Ukraine. On 7 October, a pro-war VK account praised Hamas' timing, coinciding with US artillery shipments to Ukraine.[2] Another posted that 'now all the Libtards and Ukropatriots will learn what a real Bucha looks like, and not like the theatrical performance put on by the Ukrainians.'[3]

Among Russian-language accounts on X, we observed high levels of anti-Israeli messaging from 11 October onwards. Using a customised classifier, we analysed whether messaging about the war was broadly factual, strongly anti-Israel or strongly anti-Hamas. Posts in the period 7-10 October advanced claims that Hamas was armed with NATO weapons intended for Ukraine. At this time, 62 per cent of Russian language posts from pro-Kremlin accounts were anti-Israeli. Since 11 October, it has been stable at around 80 per cent.

The calculation behind Russian propaganda appears to be this: the more the world

[1] *https://calculatebuddy.com/twitter-money-calculator*

[2] *https://vk.com/public170443089?w=wall-170443089_269296*

[3] *https://vk.com/public54012242?w=wall-54012242_2758068*

focuses on Israeli excesses, the less attention is paid to Russia's action in Ukraine. It seeks to deny the West a moral high ground and numbs the domestic audience, soothing any moral qualms about the Russian military's actions. By taking the Palestinian side, Russia is challenging those opposed to Israel's actions also to re-evaluate their positions on the Russia-Ukraine war. ■

VK, the Kremlin's Everything App

A year and a half on from Russia's invasion of Ukraine, VKontakte (VK) has emerged as the Russian everything app. Initially modelled after Facebook, VK has evolved to include functionalities akin to Apple Pay, WhatsApp, TikTok, Google Docs, YouTube, and Zoom. Post-invasion, VK launched a 100 million Rouble initiative to bolster VK Clips, introduced VK Messenger, took over Yandex News, and launched a Tinder competitor. The latter flourished after Tinder's exit from the Russian market in June 2023.

VK's growth is significant given its oversight by the FSB since at least 2014, following the forced sale and exit of its founder, Pavel Durov. Now led by Vladimir Kirienko, son of a prominent Kremlin official, VK aligns closely with government interests. In May 2023, the government transitioned over 250,000 state employees to a VK-delivered workspace. This workspace for civil servants integrated communication and productivity tools and cemented VK's role within the state machinery.

The highly anticipated ban on YouTube has yet to materialise, but during 2023, VK developed its Content Delivery Network with the intention of building sufficient capacity to replace YouTube in the event of a ban in Russia. On 4 September 2023, Forbes Russia reported widespread YouTube outages. A source attributed the downtime to actions taken by the media regulator, Roskomnadzor. The source said the regulator might be testing a mechanism for a future ban on the platform. On the same day, VK announced it was officially releasing VK Video, which would include a back-catalogue of millions of videos and exclusive new material.

During 2022, traffic to VK doubled. The company's third-quarter results, released in early November 2023, show that the platform experienced a surge in engagement, with its services reaching over 95 per cent of Internet users each month, a 16 per cent rise in daily active users to 76 million, and an 8 per cent increase in time spent on the platform. The company benefited from the onshoring of services for the education and business sectors. According to the company's reports, Educational Technologies saw a 38 per cent revenue boost, while Business Technologies, including VK WorkSpace and VK Cloud, grew by 63 per cent. ■

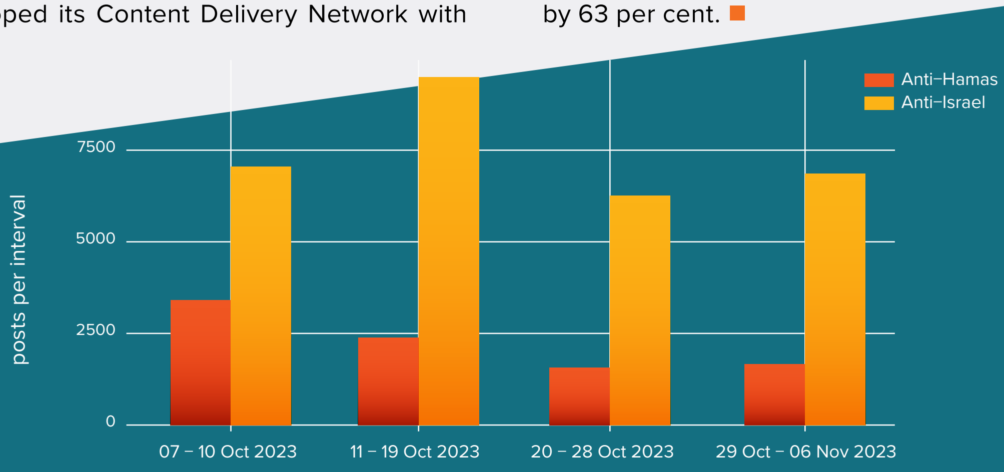


Figure 4: *Distribution of Anti-Hamas and Anti-Israel messaging from Russian-language, pro-Kremlin accounts on X*

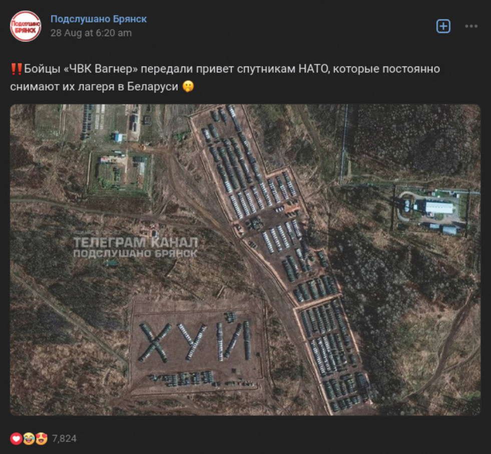
AI goes multi-modal

The latest development in AI is multi-modal models, which can handle inputs and outputs in a combination of audio, image, and text. Although open-source models such as LLaVA show remarkable results, the state-of-the-art model is OpenAI's GPT-4V, launched in late September 2023 and accessible via API from early November 2023. It can 'understand' subtext, humour, and irony in pictures, making it highly effective for meme analysis. For analysts, this means a seamless translation of visual data into text-based formats for traditional analysis tools.

Take, for instance, the content shown in Figure 5, a widely circulated VK post mentioning NATO. The vision model correctly deciphers the caption in the screenshot and contextualises it to the image. It accurately translated the caption 'Fighters of PMC Wagner send greetings to the NATO satellites constantly photographing their base in Belarus', and identified that the image shows 'an aerial view of what appears to be a military or paramilitary camp, with the characters "ХУЙ" formed using what seems to be vehicles or equipment'. It proceeded to explain that 'ХУЙ is a vulgar slang term in Russian, which is intentionally provocative. By forming this with their vehicles, the fighters are sending a defiant message to NATO, suggesting that they are aware of the surveillance and are openly mocking it.' Regarding significance for NATO, it held that 'This act showcases a brazen attitude from a group with connections to Russia. The presence of such a group in Belarus, close to NATO's eastern frontier, may be a cause for concern.'

The integration of visual models with a function calling syntax to return predictable data is particularly promising. The regular output format means API responses can be integrated into many programming workflows. For tasks that involve processing images, the algorithm can autonomously decide if trigger warnings are necessary, blurring images and providing explanations. This feature is particularly useful in conflict zones like the Israel-Hamas engagement, where graphic imagery is prevalent.

For instance, we showed the visual model a screenshot of a social media post depicting scenes from the Al Shifa hospital in Gaza. We asked, 'Is a trigger warning necessary to shield analysts?'. The response was 'yes'. 'Reason': 'The image depicts a distressing scene. It shows multiple bodies laid out on the floor, suggesting death or severe injury.'



Подслушано Брянск
28 Aug at 6:20 am

‼️Бойцы «ЧВК Вагнер» передали привет спутникам НАТО, которые постоянно снимают их лагеря в Беларуси 🤫

❤️😂😍 7,824

Figure 5: *Wagner troops send a message to NATO*

The visual model is also able to conduct relevance analysis. A simple prompt listing relevant keywords, such as NATO, security, the wards in Ukraine and Gaza, and asking, '$2s the attached picture relevant?' provides a highly accurate and reasonably efficient way to filter away material based on visual markers. We have also modified the approach to analyse videos by extracting frames at regular intervals to create a film-strip collage for assessment.

We further increase the ability to automatically determine relevance and extract summaries of visual content by including a transcript generated using a multilingual speech-to-text model. Automated image and video analysis is becoming more accessible, faster, and more cost-effective. Future issues of this brief will delve deeper into the content of TikTok and YouTube. ■

Developments in AI-Driven Manipulation

Hamas-Israel War and Fakes: AI-generated images inundated the information space during the Hamas-Israel conflict. The images, created by standard AI generators like Midjourney or Dall-E, were viewed thousands of times. They often depicted civilian suffering to garner sympathy or exaggerate support for either side, to stir patriotic sentiments. One widely circulated ***example*** shows a blood-spattered child and his dead mother surrounded by rubble. The child's raised hand has six fingers—a common AI artefact.

An ***investigation*** by NewsGuard unveiled that a fully automated fake news website fabricated the news about the death of Netanyahu's psychiatrist. The incident highlights the sophistication of AI in creating convincing yet false narratives and illustrates how fake content shared by transparently artificial sources can be laundered through social media to enter the mainstream news agenda.

Political deepfakes: The Kremlin has employed deepfakes to exacerbate internal Ukrainian tensions. One such ***clip*** depicted General Zaluzhny calling an enemy of the state and calling for a coup. Ukrainian social media users responded by creating ***humorous*** deep fakes. ***One*** showed Zaluzhny explaining there was a disagreement: Zelensky wanted to strike Sevastopol, and Zaluzhny—Kerch.

We observe a growing trend towards targeting audio rather than visual content, possibly due to the difficulty in quick and confident debunking and the scarcity of detection tools.

The Slovak election in September 2023 was ***marred*** by an audio deepfake targeting Michal Šimečka of the Progressive Slovakia party. The ***clip***, widely shared on Facebook, featured a false conversation about election rigging. Released during the pre-election media silence, it evaded timely media scrutiny. Notably, the deepfake used Slovak, a language typically difficult for automated systems to mimic, underscoring a warning to smaller language communities about their vulnerability to such attacks.

Keir Starmer, leader of the UK's opposition, was misrepresented in an ***audio deepfake*** that depicted him in a foul-mouthed rant. The clip, still circulating on X, has more than 1.5 million views. Similarly, Sadiq Khan, the Mayor of London, was subject to an audio hoax. The artificial voice criticises the Prime Minister and expresses support for pro-Palestine protests on Armistice Day.

> *"We observe a growing trend towards targeting audio rather than visual content"*

Rise of AI in Fake News Sites: Many fake news sites employ generative AI to create derivative news reports, ostensibly to gain advertising revenue. NewsGuard has ***identified*** more than 500 such sites, active in more than 15 languages. They detected the sites by searching for AI content-moderation artefacts ('As an AI model, I cannot...'). With the rise of uncensored, open-source models, such artefacts will become less common, complicating the identification of AI-generated content.

Current detection capabilities can often identify content created through proprietary models, for which extensive public training data exists. Their tendency to leave identifiable artefacts due to content moderation measures also helps. However, the emergence of uncensored, open-source models is a trickier

problem. These models can be fine-tuned to avoid replicating patterns on which detection systems are trained.

We will likely see images generated using commercial tools passed through simple filters to defeat detection systems. This can include adding various forms of digital noise, stretching, shrinking or blurring. We have already observed AI-generated fakes purportedly from Gaza that use these techniques to defeat detection systems. In one test, we found that showing a low-res version of an AI-generated image fooled a deep fake detector, reducing the stated probability of AI generation from 97 to 0.1 per cent.

As we approach important elections in Europe and the US in 2024, the landscape of digital manipulation paints a worrying picture. The ease of access to advanced audio, video, and image generation tools has opened the floodgates for misuse by various actors, from individuals to state agencies. Individuals, campaign groups, and state-level actors will inevitably experiment with these tools to achieve online information effects. ■

Prepared by Dr. Rolf Fredheim for Markolo Research 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