ISBN 978-9934-619-41-0

VIRTUAL MANIPULATION
2023
ISSUE
1

Virtual Manipulation Brief

GENERATIVE AI AND ITS IMPLICATIONS FOR SOCIAL MEDIA ANALYSIS

PREPARED AND PUBLISHED BY THE

NATO STRATEGIC COMMUNICATIONS CENTRE OF EXCELLENCE

ISBN: 978-9934-619-41-0

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

Riga, June 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

This Virtual Manipulation report explores the impact of generative AI on social media analysis. Large Language Models (LLMs), such as the powerful GPT-4, can create highly convincing content that appears legitimate and unique. This makes it nearly impossible to distinguish between real and fake accounts. But, defenders can employ the same tools to more effectively monitor social media spaces. Careless implementations by adversaries introduce weaknesses that can result in accounts inadvertently disclosing that they are bots. As LLMs rely on prompts to shape their output, targeted psychological operations ('psyops') can provoke chatbots to reveal their true identities. The fight against manipulation is entering a new phase, but it remains unclear whether, in the long run, defenders or attackers will derive greater benefit from AI systems.

At a cost of $130, we used GPT-4 to classify the content, relevance, and sentiment towards NATO for a total of 650 000 social media posts. The single event that incited the highest level of hostile anti-NATO messaging was President Putin's speech declaring mobilisation in September 2022. In contrast, Finland joining NATO in April 2023 passed with comparatively little online fuss.

In November 2022, and again in March-April 2023 the proportion of Tweets by hyperactive anonymous 'troll' accounts was ten times higher than in the first months of the war. This increase may be associated with advances in generative AI, or lax content moderation under Twitter's owner, Elon Musk.

Twitter's decision to re-amplify Russian propaganda accounts at the end of March 2023 led to the Kremlin's messaging attracting 60 per cent more views. The English language account of the Russian Ministry of Foreign Affairs saw its daily views rise from 0.44 million while de-amplified to 1.3 million per day when re-amplified.

Telegram and VKontakte have experienced consistent growth in user numbers. In March 2023, the proportion of Russians using Telegram daily exceeded that of YouTube for the first time. More than 40 per cent of Russians use these platforms on a daily basis. Instagram and Facebook, on the other hand, are accessed by around 6 and 1.5 per cent respectively. ■

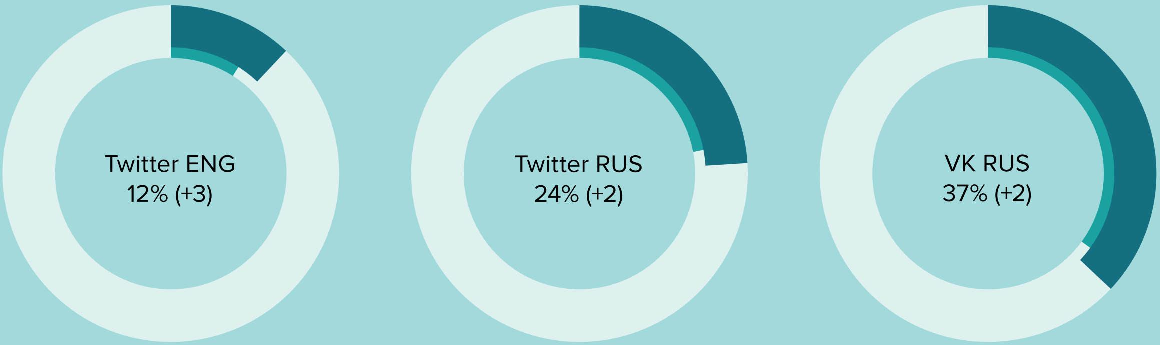

FIGURE 1: *Percentage of automated posts about NATO in Sep 2022-Apr 2023 by platform and compared to Jun-Aug 2022. In a change from our first issue, we have excluded posts in VKontakte (VK) groups, which typically make up about 50 per cent of posts.*

Russia's Evolving Russian Social Media Landscape

A year after Russia's full-scale invasion of Ukraine, the nation's social media consumption patterns have been transformed to a remarkable degree. Efforts to limit access to foreign platforms have not completely isolated the Russian internet. However, they have triggered something of a paradigm shift, prompting users to migrate between platforms. This change has become permanent and even strengthened. YouTube and WhatsApp, two of the most popular Western services, remain accessible, while platforms such as Facebook, Instagram, and Twitter are blocked. TikTok's popularity endures, though its growth trajectory has stalled as the platform has paused content creation by Russian users.

Telegram and VKontakte have seen consistent growth in user numbers. In March 2023, the proportion of Russians using Telegram daily exceeded that of YouTube for the first time. YouTube, while growing at a more measured pace, maintains its position as the most widely accessed social media platform, with 80 per cent of Russians using it at least monthly.

While Meta services Facebook and Instagram were banned in Russia as extremist, WhatsApp was exempted and permitted to continue operating. As of March 2023, over 60 per cent of the population used WhatsApp daily. Instagram and Facebook, on the other hand, have stabilised at around 6 per cent and 1.5 per cent respectively, approximately a quarter of their levels in February 2022.

The Russian government has made attempts to redirect users to domestic platforms, but with limited success. For instance, RuTube, despite having 23 million registered users and three times as many video uploads in 2022 as in 2021, ranks only as the 57th most visited site in Russia. In contrast, YouTube is ranked third.

We estimate that 24 per cent of Russian-language tweets about NATO in the Baltics and Poland came from automated accounts, compared to 12 per cent for English-language posts. This marks an increase from 22 and 9 per cent in the previous period respectively. On VKontakte, the most significant change was an increase in the proportion of posts on group pages rather than individuals' timelines, up ten percentage points from 44 to 54 per cent. Within posts to individuals' timelines, we estimate that activity from automated accounts increased by two percentage points to 37 per cent. ■

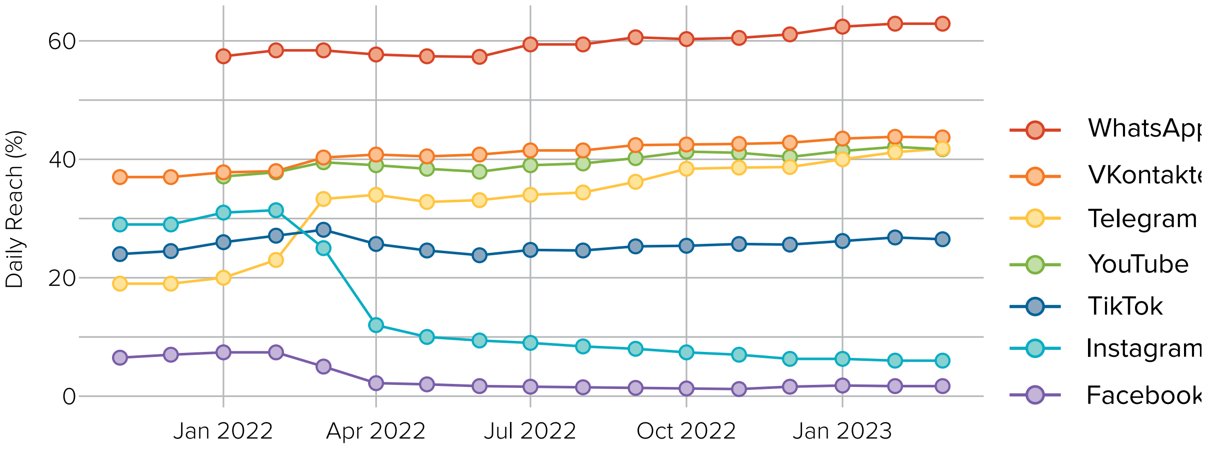

FIGURE 2: *Daily Social Media Usage by Russians (Percentage). Source: Mediascope.*

Twitter Revives Russia's Propaganda Reach

In the inaugural Virtual Manipulation Brief, we highlighted Twitter's significant strides in curbing the spread of Russian propaganda and disinformation amid the Ukrainian conflict. Measures such as 'de-amplifying' key propagandists, curtailing algorithmic reach, and suspending troll accounts tied to Russian intelligence considerably reduced the impact of pro-Kremlin messaging. Twitter's internal research suggests that these de-amplification steps cut engagement with Kremlin propaganda by 49 per cent.

However, under the ownership of Elon Musk, Twitter in late March 2023 discontinued its policy of labelling and de-amplifying Russian government and media accounts. Our analysis of eleven previously labelled accounts (Figure 3) indicates an average 60 per cent surge in views per post since policy change, corroborating findings by the DFR Lab and by Reset.

This shift led to a dramatic rise in the Kremlin's visibility on Twitter. Views per tweet for the Russian MFA's English-language account rose from 52,000 to 87,000. The policy change also fostered a more conducive environment for increased posting: the MFA's account's posting frequency more than doubled, from 8 posts per day previously to 18 per day after tha change. Comparing posts 26 days before and after shows a total view count of 34 million views, or 1.3 million daily, up from 11 million and 440,000 respectively.

Notably, Russian language accounts saw the smallest increase in views. The MFA's Russian language MID_RF witnessed a 23 per cent increase, compared to a 65 per cent rise for the English-language @mfa_russia. News outlets Izvestia and Gazeta experienced no bump in views. By contrast, the visibility of Kremlin outlets targeting an international audience such as RT Arabic and Actualidad RT saw their visibility soar by 48 and 87 per cent. Diplomatic accounts, like @RusEmbUSA, experienced a 133 per cent surge in engagement and a 150 per cent jump in views. While Twitter is becoming less relevant for Russians in Russia, it continues to be an effective tool for the Russian state to reach international and particularly non-European audiences.

We found larger mean than median increases, meaning a few highly performing tweets account for most of the boosted visibility. This underscores the effect of suppressing, and then reinstating, algorithmic recommendations. Twitter once made it significantly harder for Kremlin propaganda to go viral; now, this is no longer the case.

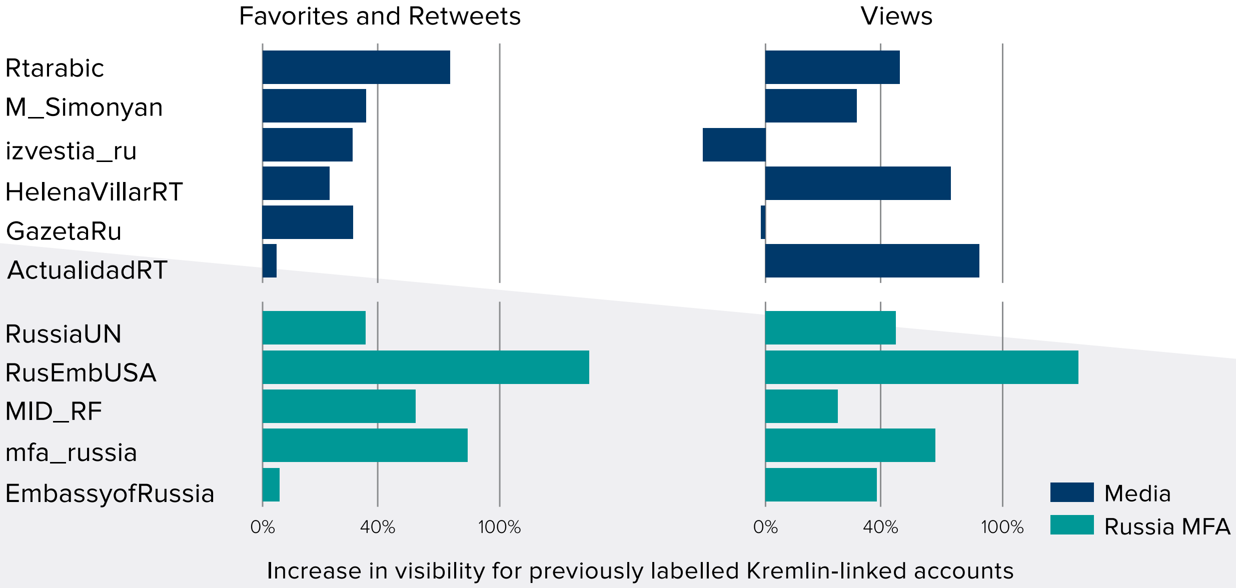

FIGURE 3: *Since March 2023, Kremlin-linked accounts have experienced a visibility boost of 60%*

Hostile messaging about NATO

This analysis examines Russian-language tweets mentioning NATO from September 2022 to April 2023. We collected 540,000 such tweets, marking a monthly increase of approximately 15 per cent compared to June-August 2022. VKontakte saw a larger rise in messaging, 25 per cent higher than the previous period, reflecting the increasing significance of the platform.

We observe a modest increase in bot activity, coinciding with Elon Musk's takeover of Twitter finalised on 27 October 2022. Although he initially promised to deal with the problem of bots on the platform, the evidence presented here shows that—if anything—the new status quo is a step backwards. However, the bigger change is by a particular type of 'troll' account: users with few followers, anonymous profiles, very high output, highly political, almost exclusively in the comments under other peoples' posts. At the start of the war, such accounts posted roughly 1 per cent of all Russian language posts mentioning NATO. In November 2022, and again in March-April 2023, they posted 10 per cent of all posts. This may indicate declining content moderation effectiveness or possible use of AI language models for creating context-appropriate, human-like posts.

High posting volume occurred in late September when President Putin cited NATO threats to justify mobilisation; in November when a stray missile landing in NATO member Poland sparked speculation about NATO's involvement in the war; and in April 2023 when Finland joined NATO.

The timeline of hostile Russian-language messaging about NATO in Figure 4 shows little criticism of Finland's NATO membership, despite past threats from Russian elites. There was a spike on 4 April, when Finland formally joined, but the number of hostile messages about weapons deliveries outnumbered those criticising Finland's decision. Instead, hostility centred on the West's perceived insincerity in ending the Donbass war, claims that NATO was supplying excessive armaments to Ukraine, and protests by Western populations against NATO and anti-Russian elites. Coverage of protests serves the purpose of conveying to Russians the sense that they are not alone.

This timeline of hostile messaging is the product of a trial using the powerful Large Language Model GPT-4 to improve our understanding of social media narratives. For a sum of $130, we made nearly 3 000 API requests, obtaining narrative, sentiment, and relevance

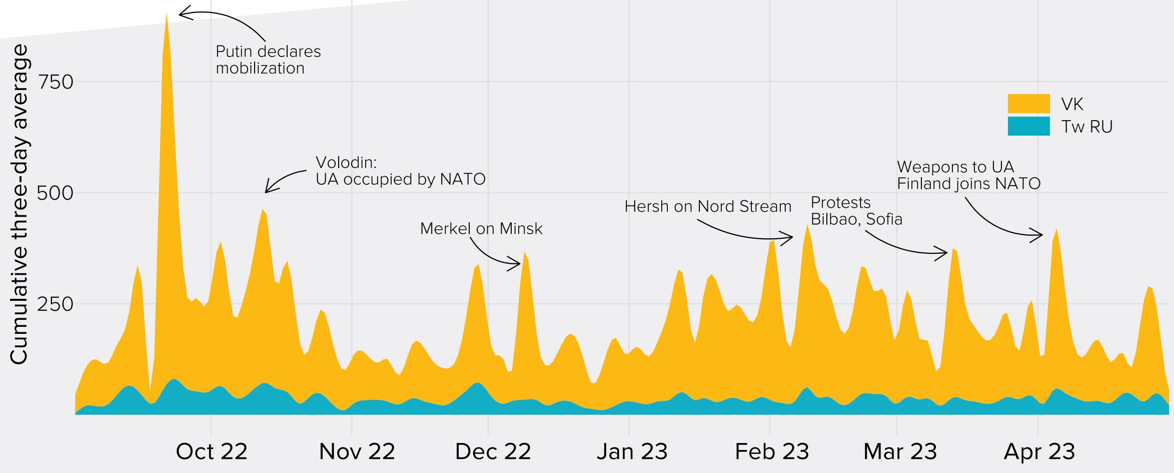

FIGURE 4: *Timeline of the most hostile Russian-language messaging about NATO*

estimates for a collection of 650 000 social media posts. We categorised textual data into discrete narratives, subsequently feeding the most prevalent ones to the GPT model. The model was instructed to assess relevance to European, NATO, and Russian politics and security, identify the antagonist, and determine the sentiment towards NATO, Russia, Ukraine, and 'the West'. Hostile messaging was determined by selecting narratives with a NATO sentiment of -5 or less on a scale from -10 to 10.

Sentiment analysis regarding a military alliance encounters a persistent problem: keyword-based methods tally instances of negative vocabulary, such as war, fight, or guns, often leading to inaccurate and negative scores. However, the AI-based approach helps circumvent this issue by focusing only on messaging about a specific entity.

Our data unlock some novel analyses: posts only tangentially relevant, such as messages about NATO allies providing aid to Turkey following the earthquakes in February 2023, can be automatically identified and excluded. Entity-specific sentiment data help identify and track messaging especially hostile towards NATO, and narratives where NATO and the West are portrayed as aggressors. This allows us to focus only on the most salient content.

Figure 5 shows who is perceived as the aggressor. Russian-language Twitter aligns more closely with English-language Twitter than with VKontakte. This observation highlights the impact of Russia's evolving social media environment since March 2022, and emphasises the effect of contrasting content moderation practices.

The large volume of material ostensibly hostile to Russia's war efforts in Ukraine is attributed to three factors: there is significant criticism of Putin's war on both Twitter and VKontakte, often posted by Ukrainians and frequently moderated, but it persists. Second, there is broad criticism from 'mil-bloggers', arguing that the Kremlin was sluggish to mobilise, criticising a perceived lack of ruthlessness, and generally viewing the military's efforts as insufficient. Lastly, the narrative sentiment tool is relatively blunt. E.g. 'Finns celebrating that the country joined NATO' is generally positive towards the alliance, but social media commentary about the subject could be either positive or negative. ■

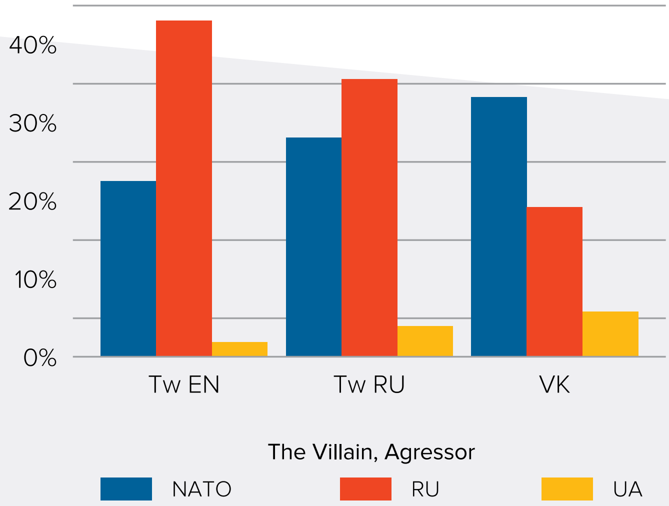

FIGURE 5: *Who is perceived as the Villain or Aggressor. The figure confirms our view that Russian language messaging on Twitter is qualitatively different to that aimed at the predominantly internal VKontakte audience.*

Social media manipulation and GPT

A new challenge for defenders

Generative AI presents challenges for social media monitoring and systems designed to detect coordinated inauthentic behaviour. GPT-4 can generate persuasive, legitimate-looking content. Unlike conventional fake accounts, posts made by top-tier language models seldom exhibit duplicates, repetitive names, or discrepancies between profile picture and name, making it challenging for traditional systems, reliant on identifying duplicated or copied content, to distinguish genuine from fabricated accounts. In the words of Alex Stamos, Facebook’s former Chief Security Officer

> *“The rapid advance of open-source generative AI is leading to a tidal wave of near-zero-cost BS flooding every text, image and video channel.”*
>
> – Alex Stamos, *April 2023*

LLMs can tailor highly personalised content to specific individuals or groups, making them ideal tools for targeting and manipulating people's opinions and beliefs. Consequently it is both increasingly important and difficult to detect coordinated inauthentic behaviour on social media. While OpenAI attempts to prevent misuse of its model, open-source alternatives evade content moderation filters, simplifying the mass production of offensive and harmful content.

The proliferation of numerous models further complicates detection. The GPT series includes major releases like 3, 3.5-turbo, and 4, all of which are continuously refined, subtly altering their behaviour. This, coupled with the continuous influx of new LLMs, results in an environment where detection systems are often outdated or optimised for an obsolete model. No wonder, a consensus is emerging that reliably detecting LLM generated text may be impossible.

Reasons for optimism

Defenders can use these tools to monitor social media spaces more effectively, and sloppy implementations by adversaries can inadvertently reveal the true identity of malicious accounts. GPT-4, if used simplistically, can fail in unusual, detectable ways.

For instance, predictable patterns emerge when content is produced at scale. When GPT is asked to generate an email, the output typically has repetitive greetings and sign-offs, as well as similar paragraph lengths. This pattern is currently easily detected as a statistical anomaly, even if individual cases may be hard to spot.

A model that typically produces human-like content may in certain edge cases generate output that restates its identity. GPT may unexpectedly change language, or refuse to create content on specific topics due to content moderation filters. In spring 2023, analysts noted an avalanche of fake accounts clearly copy-pasting GPT output by searching for messages including phrases such as “As an AI language model” or “violates OpenAI’s content policy”. Social media users have shared examples of Amazon reviews starting ‘Yes, as an AI language model, I can definitely write a positive product review about …’ Manipulators hoping to connect LLMs directly to social media accounts must prevent such content from accidentally getting posted. This is manageable, but the reality of using generative AI systems means it is not a trivial problem to overcome.

Current LLMs are surprisingly bad at generating random numbers. One researcher repeatedly asked GPT to choose random numbers between one and a hundred. In 10% of cases the model returned the number 42. This extraordinary distribution presents opportunities for uncovering fakery using statistical techniques.

There is clear potential for social media “psy-ops” targeting suspected bots: researchers can attempt to trick GPT-powered chat-bots programmed to engage with real users into revealing their targets and the viewpoints they are programmed to express. Currently, no effective countermeasures exist against prompt injection, which can seize control of AI-operated social media accounts, leading to awkward or inappropriate content. For instance, Twitter users manipulated an automated tweet bot, dedicated to remote jobs and powered by OpenAI's GPT-3, through a prompt injection attack. They redirected the bot to post absurd and compromising tweets, as well as its operating instructions. Once the exploit went viral and hundreds of people attempted it for themselves, the bot was forced to shut down.

Organised disinformation actors, scammers, and purveyors of fake news are likely excited by the possibilities of ChatGPT, but they will realise there are many reasons they cannot use these systems as part of their main operations. OpenAI maintains a centralised system, where all interactions are logged and stored under U.S. jurisdiction; systematic misuse would risk providing high-quality evidence to law enforcement agencies.

Can we use LLMs?

Sophisticated models like GPT can be incorporated in social media monitoring strategies. The section on detecting hostile messaging about NATO exemplifies this. Research indicates that LLMs excel at content analysis tasks such as classifying text sentiment or stance. However, GPT-4, as a standalone, is too costly and slow to meet the needs of extensive monitoring systems. An effective strategy is to merge the broad domain intelligence of GPT with specialised models optimised for particular tasks. This method involves deconstructing complex content analysis tasks into smaller, more manageable sub-tasks. Specialised models can then be trained based on the classifications provided by the LLM, optimising them for predictability, speed, and focus on specific tasks, such as detecting automated posting behaviour or analysing linguistic patterns. By regularly observing these specialised models' performance and using iterative loops to enhance their functionality, the system can adapt to new tactics used by malicious actors while maintaining accuracy and consistency, and minimising the risk of hallucination or unexpected outputs.

> *"Leveraging generative AI in automated systems is hard; manipulators will struggle to operate both securely and at scale."*

LLMs can be trained to perform repetitive tasks normally conducted by human analysts, such as scanning a user's timeline to identify new trends or themes, inconsistencies in their expression, or evaluating the likelihood of a timeline being automated. In each case, the human analyst can assess the accuracy of the model's output, and these assessments in turn provide valuable training data that can be used to further increase accuracy.

Furthermore, LLMs could perform routine tasks, allowing analysts to concentrate on core duties. One example is an automated scanning and reporting system that examines logs for errors or gaps in data collection. Or a more robust implementation where the LLM compares aggregate metrics for a given monitoring period to those of the previous one, identifying anomalous or unusual patterns and accounts, and summarising this into an auto-generated report. This AI-driven reporting system could be further customised for different target audiences, e.g. highlighting bugs and errors to the developer, material for social media posts to the PR officer, and key takeaways to the decision maker.

Considerable hurdles prevent LLMs being adopted by military and governmental institutions. These include risks associated with handling sensitive or classified documents, as well as personal information and various privacy concerns. Sending an institution's code-base or internal documents to a cloud-service is not an option. However, these challenges can be mitigated through a combination of in-house models, stringent filtering and encryption strategies, and collaboration with trusted and vetted vendors. Now is the perfect time to experiment with tools such as Nomic's GPT4ALL and Private GPT—both of which enable powerful LLMs to be run on local servers. The potential of LLM-assisted social media analysis is immense; defenders ignore it at their peril. ■

About the Virtual Manipulation Brief

The Virtual Manipulation Brief carries forward and expands the remit of NATO StratCom COE's Robotrolling series. 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 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.

Prepared by 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