ISBN 978-9934-619-07-6

Social Media Manipulation for Sale

2024 Experiment on Platform Capabilities to Detect and Counter Inauthentic Social Media Engagement

PREPARED AND PUBLISHED BY THE

NATO STRATEGIC COMMUNICATIONS CENTRE OF EXCELLENCE

ISBN: 978-9934-619-07-6

Authors: Dr Gundars Bergmanis-Korāts, Tetiana Haiduchyk
Project Manager: Dr Gundars Bergmanis-Korāts
Contributors: Bohdan Smolts, Cyber Department of the Security Service of Ukraine
Experiment: Trementum Research
Content Editor: Egil Fredheim
Design: Inga Ropša

DISCLAIMER
This report was completed in November 2024, based on an experiment that was conducted in August and September 2024. Discussions with social media companies regarding preliminary results took place in October 2024.

Riga, November 2024

NATO STRATCOM COE
11b Kalnciema iela,
Riga, LV1048, Latvia
stratcomcoe.org
@stratcomcoe

Social Media Manipulation for Sale

2024 Experiment on Platform Capabilities to Detect and Counter Inauthentic Social Media Engagement

Content

Executive Summary

Two and a half years since the Russian invasion in Ukraine, in addition to newly adopted EU regulations such as the Digital Services Act (DSA)[1] and the Digital Market Act (DMA)[2] allowed us to continue the series of these reports with high expectations of platforms having developed capabilities to identify and remove commercial manipulations. Although there were minor improvements across most platforms since our last experiment in 2022, our expectations were not fully met, leading us to speculate that EU regulations had minimal impact on detecting inauthentic engagement during our red team experiment. Why was this the case? Several potential reasons, which we explore in detail throughout this report, may address this phenomenon.

During the experiment where we purchased inauthentic engagement from commercial social media manipulation services, platforms demonstrated significant variation in resilience to inauthentic activity, with notable differences in the ease of registration and the cost of SMS verification.

Most platforms struggled with the removal of fake accounts; X showed good progress by removing 50% of identified fake accounts, while TikTok and VKontakte managed to remove only a small fraction (3% and 2%, respectively). In addition, fake interactions remained prevalent across all platforms, highlighting ongoing challenges. Manipulation services have become increasingly affordable, even from reputable providers in the US and UK, making inauthentic engagement more accessible. While the majority of clients using manipulation services are commercial entities seeking to promote spam, scams, or other commercial topics, we observed the use of bots to amplify political content on a diversity of topics. In the context of the US political environment, we observed bots being utilised to influence public opinion by amplifying divisive content related to the upcoming elections. Bots engaged in promoting and countering narratives about prominent political figures, such as President Joe Biden, Kamala Harris, and Donald Trump, as well as other politically charged content. This trend highlights the ongoing vulnerability of platforms to manipulation in politically sensitive environments, despite efforts to curb inauthentic engagement. The platforms' efforts to counteract various forms of inauthentic activity, such as fake likes and views, have largely been ineffective, with X standing out as the only platform making significant progress by removing approximately 50% of fake comments and reposts.

Most platforms showed a considerable interest in our findings, and many provided insights into their transparency reporting. We reached out to all platforms except VKontakte and received responses from all but X. A key conclusion from our experiment is that, from a red-team perspective, it was successful due to the complex content and behaviour classification systems employed by platforms. These systems rely on multiple indicators to determine whether content should be flagged as problematic, yet our experiment likely remained undetected due to its small scale. X demonstrated that small-scale commercial manipulation can be identified and removed effectively, whereas other platforms continue to struggle with this challenge. A concerning takeaway is that malicious actors can evade detection by breaking large-scale campaigns into smaller micro-scale operations, using different commercial services, and interacting sporadically with platforms. This ability to remain undetected presents a significant threat to the integrity of social media ecosystems.

Introduction

In this fifth iteration of our social media evaluation, which we have been conducting since 2019, our core objective remains the testing and assessment of social media platforms' resilience to manipulation by well-funded commercial manipulation service providers. In this report, we outline the methodology of the experiment and present the results for each assessment criterion. The experiment is designed to evaluate the platforms' ability to detect and remove commercial manipulation, specifically in non-political contexts. This enables us to identify and track the accounts involved in such manipulation and assess their activities across different platforms. Finally, we conclude the report by offering insights into the operations of bot farms and their role in supporting disinformation campaigns targeting Ukraine.

In previous periods, we observed progress in social media companies' efforts to combat manipulation on their platforms. However, our findings revealed that during 2022/2023, the platforms' overall capacity to prevent manipulation declined. While certain platforms made minor improvements in specific areas, none demonstrated a comprehensive improvement compared to the 2021 report. Despite the ongoing Russian war, social media manipulation remained both inexpensive and easily accessible. Moreover, a significant portion of the purchased inauthentic engagement was not removed by the platforms, even after being reported. These findings indicate that the efforts of social media companies to counter manipulation may be insufficient, highlighting the potential need for more effective regulation. With the decline in platform effectiveness in the previous period, the key question is: has the situation improved? Have social media companies made any significant progress in combating manipulation, or do the same challenges persist?

The Experiment

Methodology

During August and September 2024, we conducted an experiment to test the ability of social media companies to identify and remove manipulation. Within the scope of this experiment, we restricted our use of commercial manipulation services (purchased inauthentic engagement) to non-political contexts. This approach allows us to assess the platforms' ability to detect commercial manipulation (fake engagement delivered by bots). We purchased engagement (likes, views, shares & comments) on 44 fake posts we have created using fake accounts we registered, enabling us to apply our assessment criteria, which include indicators such as account blocking, delivery speed, the remaining share of accounts and engagement, as well as the responsiveness and transparency of company reporting.

For €58, we received 1,150 inauthentic comments, 11,725 likes, 3,150 shares, and 8,233 views on Facebook, Instagram, YouTube, TikTok, VKontakte, and X. For approximately the same amount of engagements, it is 3 times less than in the previous assessment period.

Improvements

This year, we expanded the number of commercial manipulation providers to six, including two from the US, one from the UK, one from Italy, and two from Russia. We retained the assessment criteria from the previous *report* while broadening the evaluation of platforms' ability to block the creation of inauthentic accounts. This was achieved by introducing a new indicator that compares SMS verification costs across various countries.

Assessment Criteria

In this report, we will assess social media platforms based on several key criteria. These include their effectiveness in preventing the creation of inauthentic accounts, as well as their ability to detect and remove both such accounts and the inauthentic activities they generate. We will also evaluate the cost of manipulation services, examining the affordability and accessibility of these services across different regions. Furthermore, we will assess the speed and prevalence of manipulation activities, the platforms' responsiveness to these threats, and the transparency of their actions and reporting, which will all be essential aspects of our evaluation.

1. Blocking the creation of inauthentic accounts

The experiment showed that registering accounts on most platforms was relatively straightforward. Facebook, Instagram, X, and YouTube presented no significant obstacles during the account creation process, but some platforms presented technical challenges. TikTok had issues with phone verification on its web version, requiring multiple attempts to receive the confirmation code. VKontakte proved to be the most challenging, as registration was only possible through their mobile app, which was inaccessible in certain regions, necessitating a workaround to download the installation file directly from their website. In addition to previous assessments, this year we decided to improve and broaden these particular assessment criteria by adding a comparison of SMS verification cost[3] where we analysed the prices and traffic of fake SMS verifications. The number of countries in which we observed such services differs depending on the platform: Instagram, X and VKontakte—177; Facebook—182; TikTok—175; YouTube—168.

A notable point is that YouTube's average price emerges as the highest among the platforms in this comparison, and therefore, as we expanded the criteria, we decided to add this as an improvement in the overall assessment. In contrast, TikTok offers the lowest price for verification, while Facebook displays the highest observed cost.

So what? The differences in registration ease and SMS verification costs point to varying vulnerabilities across platforms. Higher SMS costs, such as on YouTube, may signal stronger defences, while TikTok's lower costs could indicate weaker safeguards. The low cost of inauthentic accounts allows malicious actors to automatically or with low effort create large volumes of fake accounts, enabling them to swiftly scale their campaigns.

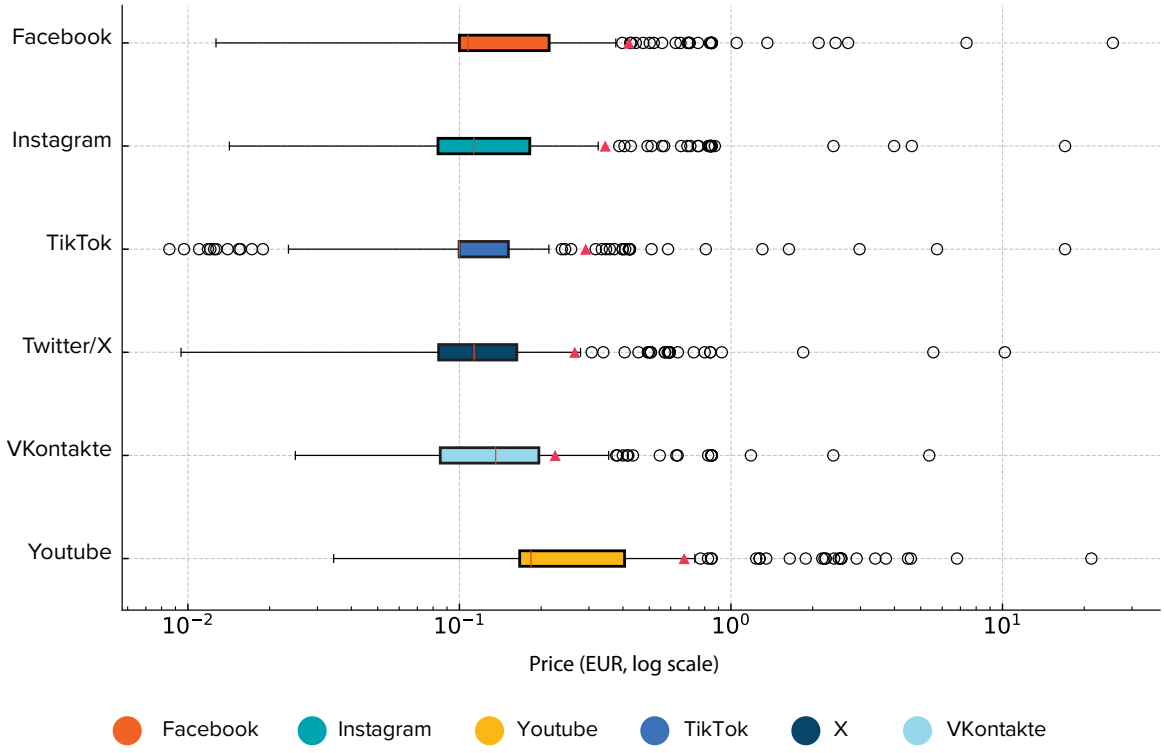


FIGURE 1. Price range for SMS verification on platforms

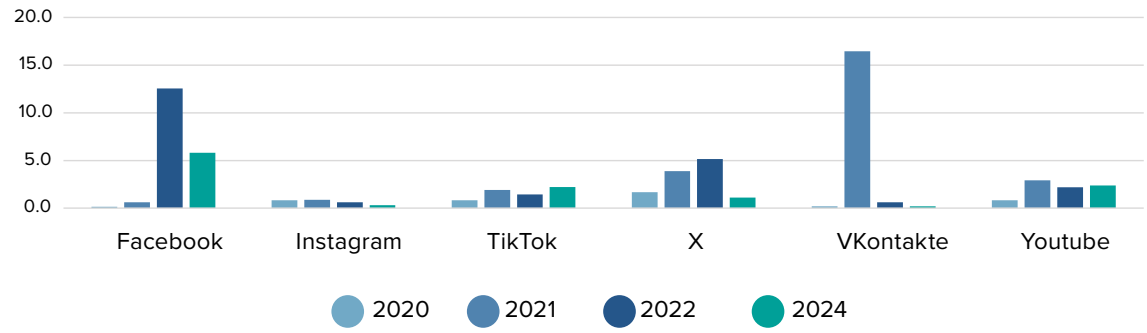


FIGURE 2. Cost of inauthentic accounts

Since 2022, the cost of inauthentic accounts has generally declined, except for those on TikTok and YouTube. The most substantial price drop occurred with X accounts, which are now more than four times cheaper. This assessment is based on the minimum prices offered by various online wholesalers of fake accounts.

2. Removing Inauthentic Accounts

This time, an average of only 15% of identified accounts were removed over the period of four weeks, which is the lowest rate over the last two experiments (25% in 2021 and 16% in 2022).

However, this year, VKontakte saw a significant drop in effectiveness, with only 2% removed, the lowest rate recorded. As in the previous results, fake engagement remains when accounts are removed.

In contrast, **X showed the best results, with 50% of removed fake accounts during the monitoring period**. Facebook, compared

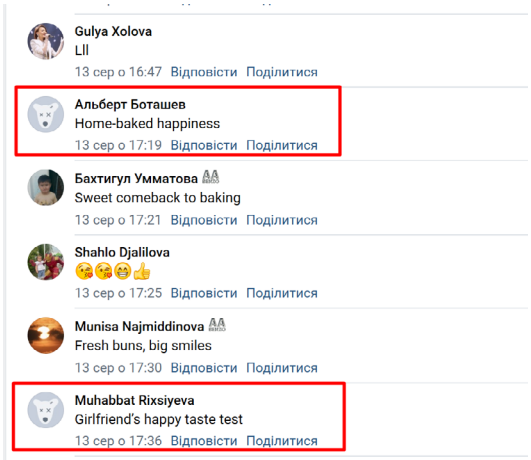


IMAGE 1. An example of inauthentic accounts

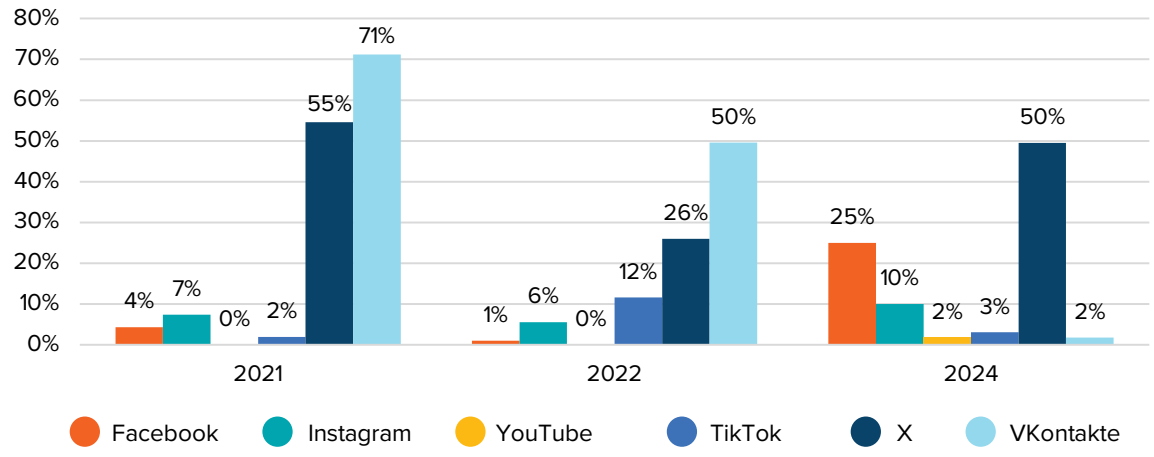


FIGURE 3. Removed inauthentic accounts during the monitoring period

to previous experiments, also demonstrated a relatively high performance, with 25% of accounts removed. Instagram showed stable results relative to previous studies, but TikTok significantly worsened its performance, with only 3% removed. YouTube, while improving its performance, still only managed to remove 2%.

So what? The low removal rates of identified fake accounts across most platforms indicate a persistent vulnerability to manipulation, with X showing the best improvement (50% removed) while others like TikTok (3%) and VKontakte (2%) performed poorly. This lack of consistency highlights the fact that many of the platforms still have trouble tackling the blocking and removal of inauthentic accounts, and thus face an ongoing struggle to maintain information integrity..

3. Removing Inauthentic Activity

In this experiment, we found that platforms such as X and YouTube showed significantly better results compared to the previous research. X demonstrated the best performance, with 61.91% of inauthentic engagement remaining. For one of the tweets, retweets were completely removed within 4 weeks. Facebook and Instagram also showed slightly better results than the previous experiment. VKontakte remained at the same level as in the previous research, and TikTok performed the worst. Overall, 86% of inauthentic engagement remained active 4 weeks after purchase.

So what? While X and YouTube showed improvements, with X performing best, a considerable amount of fake engagement still remained across platforms after four weeks. TikTok fared the worst, and Facebook and Instagram saw only slight improvements. These findings highlight the ongoing need for stronger efforts to tackle the identification and removal of inauthentic engagement. The prompt removal of inauthentic activity significantly impacts how much these interactions can influence discussions on social media, making it a vital responsibility for platforms.

	2020	2021	2022	2024
Facebook	96.53%	98.52%	99.49%	93.48%
X	74.23%	83.43%	82.27%	61.91%
Instagram	91.80%	96.01%	99.94%	98.62%
TikTok	99.69%	84.77%	97.33%	99.85%
VKontakte	-	99.96%	99.92%	99.32%
YouTube	97.17%	92.38%	90.00%	78.71%

TABLE 1. Percentage of inauthentic activity remaining on the platforms after four weeks

4. Cost of Services

We compared the price of a **basket** of manipulation consisting of 100 likes, 100 comments, 100 followers, and 1,000 views from six Russian manipulation service providers to arrive at a median price for 2024. We compared it to assessments of previous years and historical data.

Since 2022, the price of manipulations has decreased for all platforms except VKontakte, which experienced a slight increase. This year, the most significant price drop was observed for Facebook. In 2024, manipulation baskets for YouTube and TikTok have also seen price reductions, while

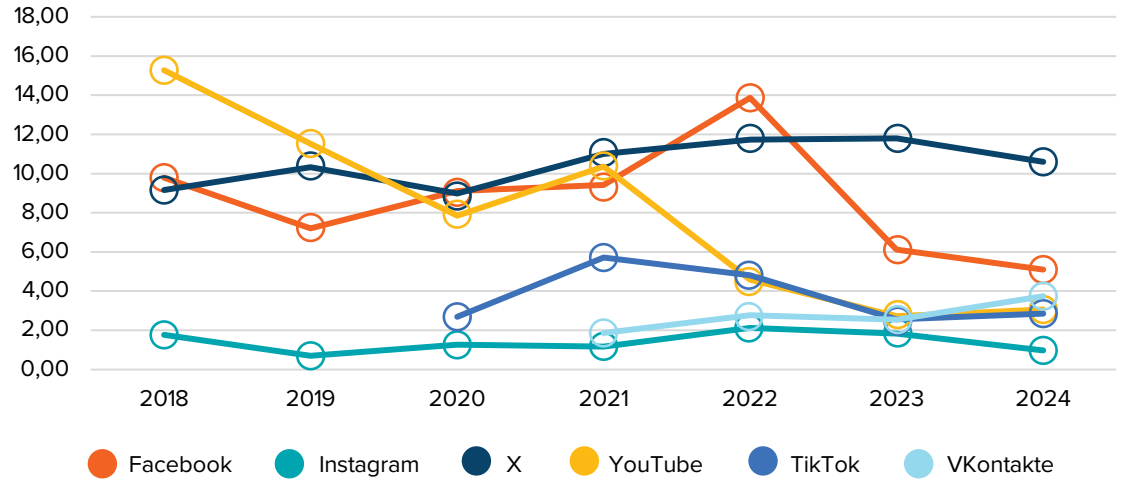


FIGURE 4. Price of a basket of social media manipulation

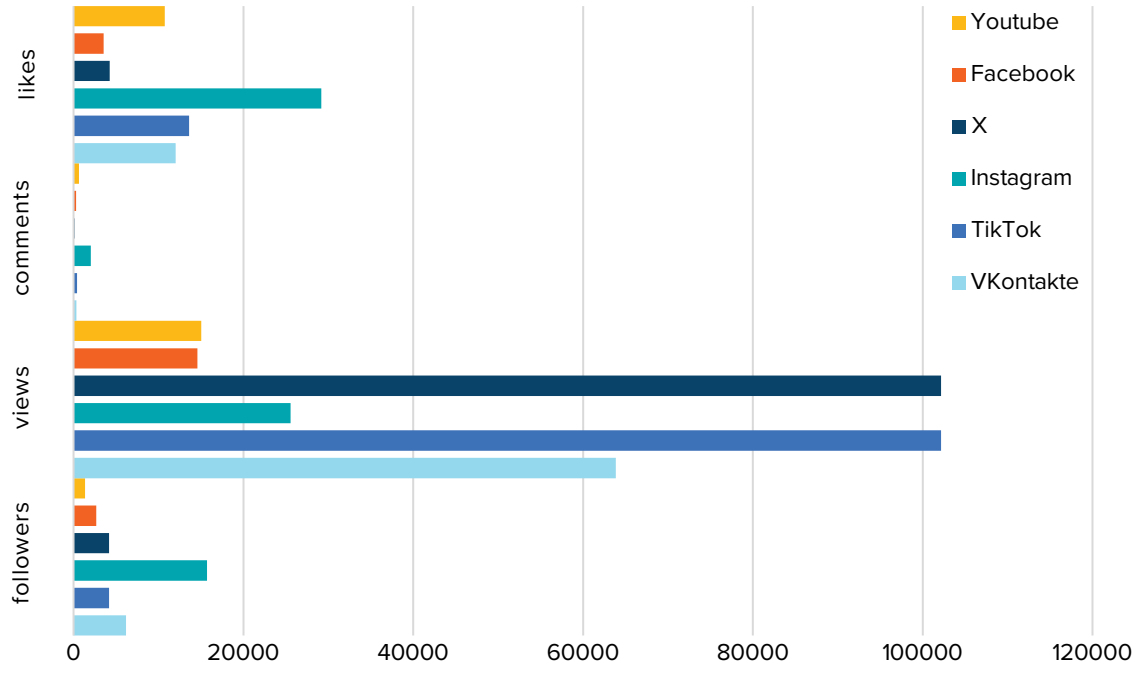


FIGURE 5. How much manipulation can you buy for 10 EURO?

Instagram remains the cheapest among all social networks.[4]

Labor-intensive manipulations, such as comments, remain several times more expensive than automatic manipulations, such as views or likes. For €10, you can buy a significant number of non-authentic views on X and TikTok, making it one of the most profitable platforms for viewing. In contrast, Instagram offers the highest number of likes for the same price, while platforms like YouTube and Facebook are the most expensive in terms of followers. There you can buy significantly fewer inauthentic followers compared to other platforms.

Comparing providers' prices

Contrary to expectations, the most expensive provider is based in Russia, whereas providers from the US and the UK, which we utilized this year, proved to be the cheapest. However, some of the assessed Russian providers lacked certain services, resulting in lower prices for platforms like Facebook, X, and TikTok due to these omissions. Nevertheless, these more affordable providers still offered the lowest overall prices. Meanwhile, a provider from Italy was slightly more expensive than average.

So what? Manipulation services are becoming more affordable on key platforms like Facebook, YouTube, and TikTok, making it easier and cheaper to spread inauthentic engagement. This trend, combined with the surprising affordability of services from US and UK providers, highlights the growing accessibility of online manipulation, posing a continued threat to the integrity of online platforms.

5. Speed and Availability of Manipulation

In 2024, we tracked that 93% of all engagement across all platforms was delivered within the first 24 hours. Comparing this figure to the 2022 investigation, which reached 89%, we can see that the speed has increased this year.

YouTube and TikTok showed the worst results, delivering more than 100% of the expected engagement within the first 12 hours. **Whereas X demonstrated the best performance, with 64% of the expected engagement achieved within the first 24 hours.** This was the highest percentage recorded during the entire monitoring period, after which the engagement rate only declined.

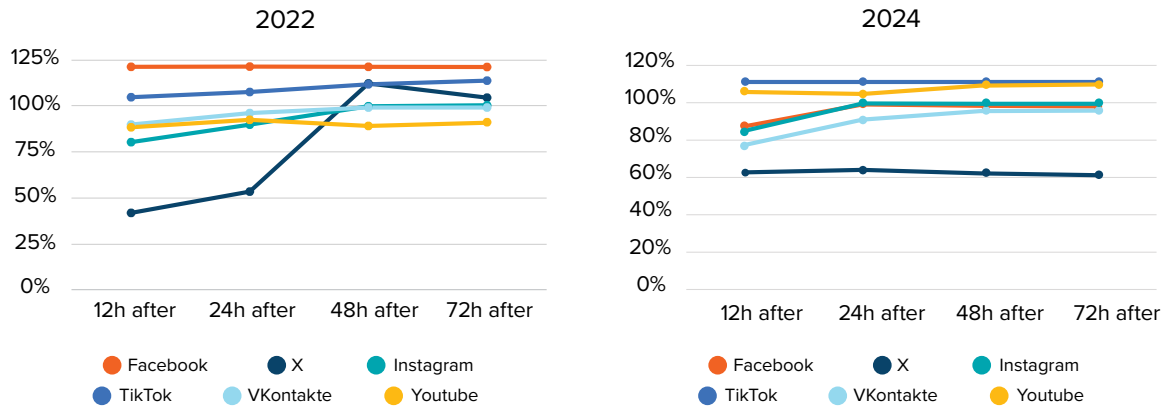


FIGURE 6. Compared to 2022, the overall speed delivery seems to be changed for all platforms except VKontakte, which may indicate a lack of changes on the platform

So what? Faster delivery of inauthentic engagement makes manipulation harder to detect and counter in real-time. This growing efficiency, especially on platforms like YouTube and TikTok, poses a significant threat to the integrity of online interactions, as platforms may struggle to address manipulative behaviour quickly enough to mitigate its impact.

6. Responsiveness

We reviewed the reported accounts in the 2022 study for six platforms, and none of them succeeded in fully addressing the issue of inauthentic accounts that may be involved in social media manipulation. Platforms like YouTube and TikTok performed the worst, with 82% to 95% of reported bots remaining active. VKontakte, Facebook, and Instagram had average results, with 54% to 67% of bots still active. However, it's worth noting that X saw a sharp decline in bot accessibility, with only 11.3% remaining active.

After monitoring the accounts for five days, we found that Facebook showed the best results, with a 6% removal rate on the first day. VKontakte followed behind, with over 3.3% removed. X, Instagram, and YouTube delivered average results, with removals ranging from 1.3% to 2.7%. In contrast, TikTok performed the worst, with a removal rate of 0.7%.

A small improvement when compared to 2022 among all platforms except TikTok in removing accounts five days after reporting.

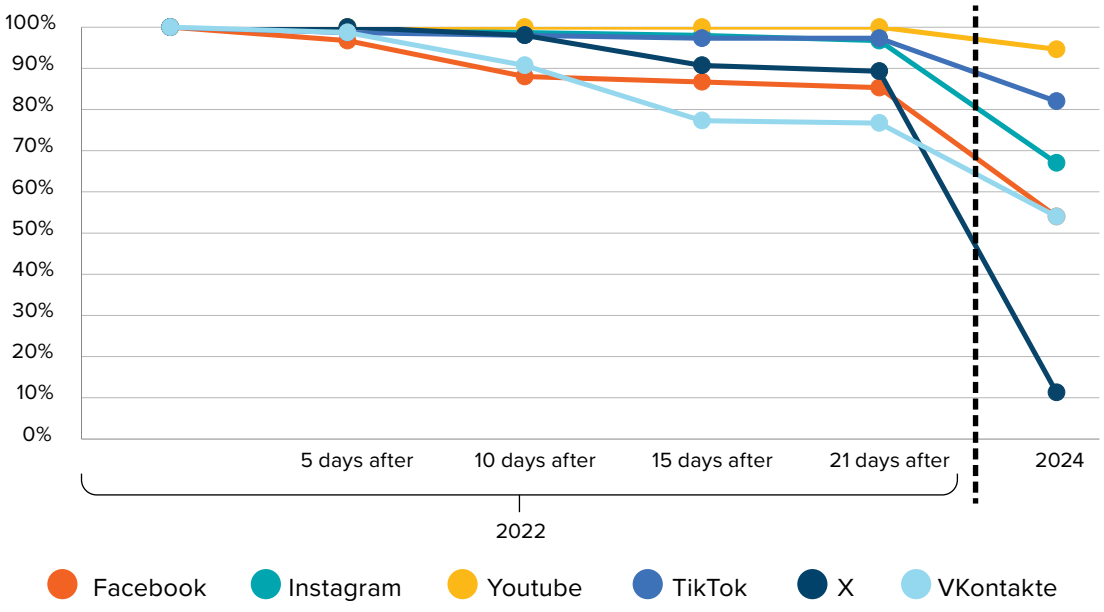


FIGURE 7. Share of reported accounts remaining active after reporting in 2022

This year, in the first five days after reporting, the removal rate of inauthentic accounts ranged from 0% to 6%, indicating a slightly better result than the previous experiment.

So what? Most platforms remain ineffective at removing inauthentic accounts, with only slight improvements since 2022. The continued presence of a large percentage of bots, especially on platforms like YouTube and TikTok, means that manipulation remains

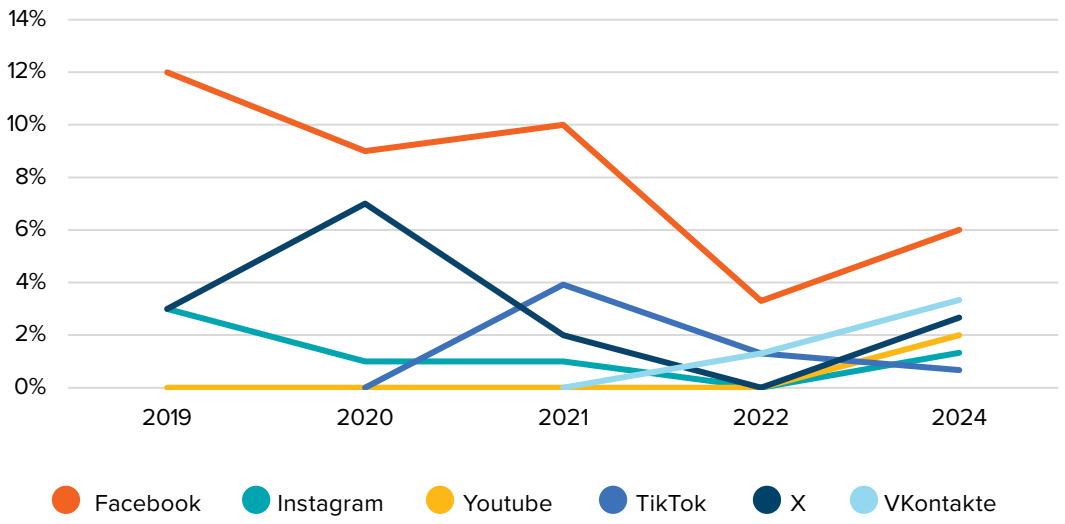


FIGURE 8. Share of accounts removed five days after reporting

a serious and ongoing issue. While X showed progress, the slow removal rates across platforms signal that current measures are still inadequate to protect against manipulation in a timely manner.

7. Transparency of Actions

TikTok, *X*, *Meta*, *Youtube*, and *VKontakte* all provide transparency reports, but the depth and focus of these reports vary. TikTok clearly reports on fake account removals when YouTube and X report the number of accounts removed for being dedicated to spam. X, in turn, provided the first update of the transparency report since 2021. Meta continues its practice of only providing transparent data regarding the taking down of fake accounts for Facebook while failing to do the same for Instagram.

We reached out to all platforms (except VKontakte) and shared the main results of our experiment. We inquired about their perspectives regarding the reasons for the failure of our assessment criteria. While we received responses from all platforms except X, only Meta and TikTok demonstrated interest in discussing our findings and provided answers to our questions. In contrast, Google requested additional information regarding the experiment but did not offer their perspective.

So what? Inconsistent transparency reporting from major platforms hampers our ability to fully understand and combat inauthentic activity. Without detailed, uniform data, platforms can avoid accountability, leaving researchers and policymakers with an incomplete view of manipulation, making it harder to counter. Platforms face technical challenges, such as scaling and language barriers, and likely withhold detection details to avoid “educated” manipulations. Our experiment shows that small-scale inauthentic manipulations often go undetected, allowing malicious actors to run dark PR campaigns undetected by sophisticated machine-learning enabled detection systems.

Recycled puppets

This section analyses how inauthentic accounts, initially used in the experiment, are later repurposed. Researchers tracked 6,632 such accounts, examining their posts and follower patterns. While promoting personal blogs and cryptocurrency was common across platforms, some activities were platform-specific. For instance, bots on X were observed publishing pornographic content. Notably, the study found a concerning trend: the increasing diversity of political content promoted by these accounts. While most clients using manipulation services are commercial, the use of bots to amplify political content is a growing concern, indicating potential manipulation of elections and political processes. On X, we identified around 17 topics bots engaged with, 4 on VKontakte, 3 on TikTok, 1 on Facebook and none on Instagram and YouTube.

This section outlines how the consumers of the social media manipulation market reuse the inauthentic accounts that provided manipulation services during our experiment. It is important to note that, due to insufficient data from the platforms, we were unable to identify the accounts responsible for specific inauthentic activities. For instance, we could not trace the accounts delivering fake shares on Facebook and VKontakte, while on YouTube and TikTok, we were only able to track the accounts involved in posting inauthentic comments.

For this iteration of the report, we managed to identify and track the activity of 6,632 inauthentic accounts used by manipulation providers to promote posts from our experiment. Under bot activity, we mean not only the content they post but also an analysis of their followings.

FIGURE 9. Tracked inauthentic accounts by delivered engagement type

We observed several similarities in the use of commercial bot accounts across social networks. For example, promoting personal blogs was common for all platforms, while crypto-related accounts were amplified by bot activity on X, TikTok, and VKontakte. On Instagram and VKontakte, bots also frequently interacted with online stores offering a variety of products.

Regarding the differences, we recorded bots interacting with YouTube channels that published game content. Whereas we only detected a commercial bot publishing its own pornographic content on X, we tracked the interaction of almost 200 bot accounts with the Onlyfans model account on Instagram.

As for political content, potential consumers of manipulation services vary between platforms. Although we didn't observe any commercial bots interacting with political content on YouTube and Instagram and just a few cases on Facebook during this iteration of the experiment (in 2022 Facebook had notable amplification of Azerbaijani governmental content, and YouTube bots were heavily involved in boosting pro-Kremlin channels), we detected numerous examples on TikTok, VKontakte, and X.

On VKontakte, bots continue to promote pro-Kremlin content that justifies the war against Ukraine, criticises the West, and calls for mobilisation into the Russian army. Additionally, as observed in the previous study, we recorded interactions of bots with posts promoting the Liberal Democratic Party of Russia. Overall, bot accounts were utilised to disseminate content targeting a Russian domestic audience on VKontakte.

Despite TikTok remaining partially accessible in Russia, we identified numerous instances of inauthentic users promoting content targeting the Russian audience. In particular, the bot accounts interacted with an account identifying itself as a 'Russian Occupant' and supporting Wagner PMC. Other examples included praising life in Russia, celebrating National Flag Day, and justifying the war against Ukraine.

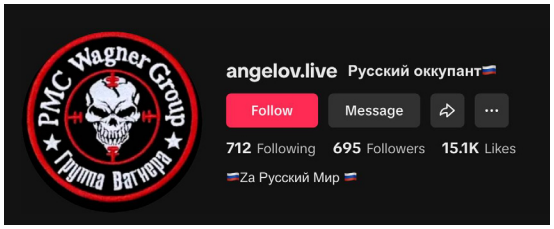


IMAGE 2. TikTok account "Russian Occupant" supporting Wagner PMC

However, in contrast to VKontakte, commercial bots on TikTok were also used to promote content to other audiences. In particular, we recorded the interaction of inauthentic accounts with content that criticised the Moldovan government and President Sandu regarding Moldova's pro-European orientation and supported the pro-Russian presidential candidate, Irina Vlah. Additionally, manipulation services were utilised to disseminate content targeting the Ukrainian audience through anti-mobilisation narratives.

https://www.tiktok.com/@treshkas3/video/7384517633799982341

IMAGE 3. TikTok video with fake likes

The most diverse range of political content with which bot accounts interacted was observed on X. Similarly to TikTok and VKontakte, we observed Pro-Kremlin content amplified by commercial bots on X. Specifically, bots from our sample reposted content as if coming from official accounts, such as the Russian Embassy in the UK and so-called Kremlin 'talking heads' like Ukrainian journalist Diana Panchenko, who worked at the pro-Russian TV channel owned by Viktor Medvedchuk, and pro-Russian Irish journalist Chay Bowes. Another example of the bots' activity in promoting pro-Kremlin content is the reposting of tweets that cite Maria Zakharova, the spokeswoman for the Ministry of Foreign Affairs of the Russian Federation.

Another interaction of bots was observed with pro-Chinese accounts discussing the relationship between the USA and Taiwan and condemning Taiwan's independence.

Furthermore, X commercial bots from our sample were involved in promoting content related to the US political environment. The bots interacted with tweets about local politicians, the current president, Joe Biden, and candidates for the upcoming elections.

https://www.tiktok.com/@treshkas3/video/7384377241871846662

IMAGE 4. Pro-Kremlin TikTok video with fake likes


迷人的小红
@miren_41319

从立陶宛到什么圣文森岛到美国跨党派人士……这一系列恶心的窜台小动作，都直指美国一个险恶用心，就是想把中国拉入战争漩涡！
一来可以卖军火，二来可以让中国内斗元气大伤，三来为它日薄西山的地球霸主垂死挣扎一下。最主要是，利用这些不入流的伎俩，拖住甚至破坏中国发展的节奏和脚步。

Translated from Chinese by Google

From Lithuania to St. Vincent Island to bipartisan figures in the United States...this series of disgusting little actions in Taiwan all point directly to the sinister intention of the United States, which is to drag China into the vortex of war!
First, they can sell arms, second, they can cause China to suffer from internal strife, and third, they can make a last-ditch effort for their declining status as the world's hegemon. Most importantly, they can use these inferior tricks to slow down or even disrupt the rhythm and pace of China's development.

Was this translation accurate? Give us feedback so we can improve:

5:14 PM · Aug 14, 2022

34 21 224

https://x.com/miren_41319/status/1558819456788332544

https://x.com/SpokespersonCHN/status/1794014561294819660

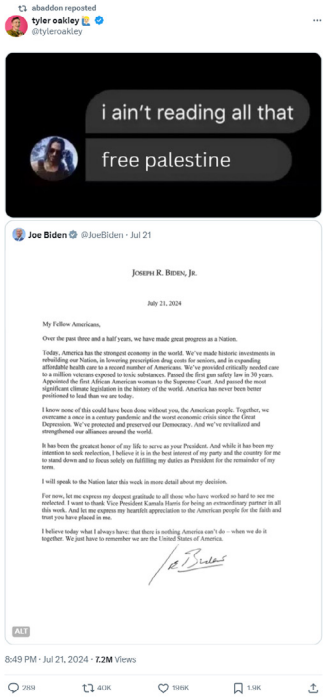
IMAGE 5., IMAGE 6. Tweets reposted by commercial bots on X

Specifically, we recorded the bots' support in the replies to tweets criticising the Democratic Party, President Joe Biden, and presidential candidate Kamala Harris. Bots generated replies to posts by Trump supporters, amplifying a post announcing a "GROYER WAR II" against the GOP and Trump's campaign and interacting with accounts featuring pro-Trump banners and hashtags like "MAGA" or "MAHA". Bots were also utilised to counter negative sentiment towards influential figures like Elon Musk. We detected bot replies from our dataset to posts supporting Trump's presidential campaign. We tracked similar activity on Facebook, where over 100 bots followed an account that called for voting for Trump.

https://x.com/southernsass81

https://x.com/HappyCJ23

IMAGE 7., IMAGE 8. Examples of X profiles whose posts were commented on by bots



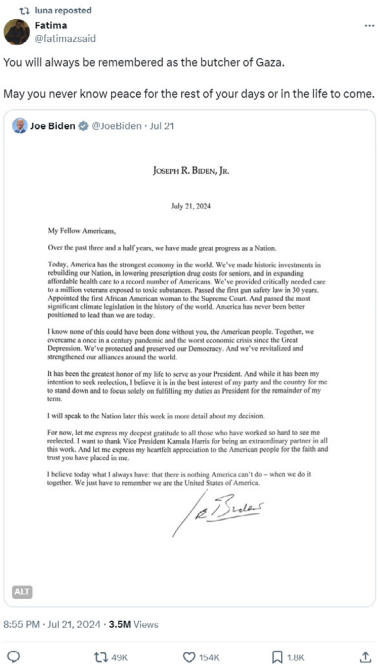
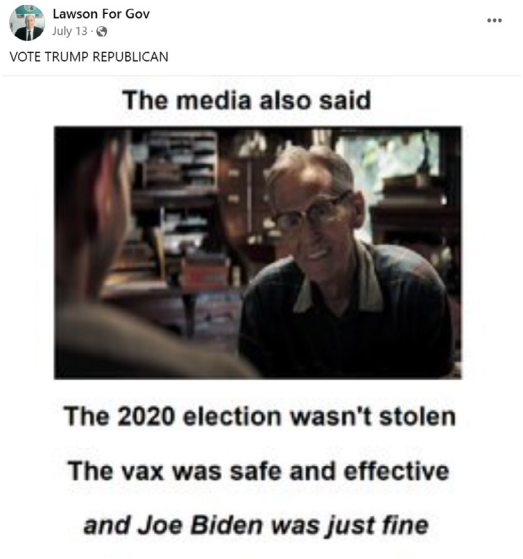


IMAGE 9., IMAGE 10. Examples of X posts reposted by bots



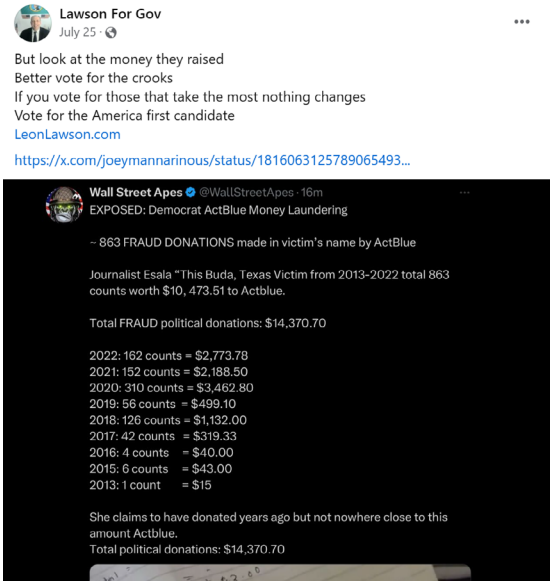


IMAGE 11., IMAGE 12. US-related content on Facebook account followed by bots

Although most US-related content on X amplified by bots had anti-Democratic Party sentiment, we also found one instance where inauthentic accounts left favorable comments on a tweet expressing support for the Democratic Party and presidential candidate Kamala Harris. Additionally, we also observed bot activity in disseminating content related to the EU political environment. In particular, inauthentic accounts amplified tweets criticizing Thierry Breton, the European Commissioner, who emphasized in his tweet the importance of adhering to the Digital Services Act and claimed to have sent a letter to Elon Musk to discuss these issues.

Another instance of bot interactions was recorded on tweets about the Olympic Games in Paris. Inauthentic accounts supported content criticizing the World Anti-Doping Agency, accusing it of manipulating in favor of Western countries, and amplified

IMAGE 13., IMAGE 14. Content on X amplified by bots

a post misleadingly discussing the gender of Algerian boxer Imane Kheli. Other groups of bots from our sample posted negative comments on tweets supporting LGBTQ+ activist.

So what? As in previous iterations, this study confirms that the majority of clients utilizing social media manipulation services are commercial entities. **However, a concerning trend is the increasing diversity of political content promoted by commercial bot accounts each year.** This indicates that influence operations related to elections and other political processes can be readily amplified.

Engagement focus - what matters and where?

We recognise that different forms of engagement incur varying costs, with the understanding that commenting is generally more complex and resource-intensive than likes or views. Comments serve not only as a quantitative measure of engagement but also encapsulate sentiment and opinion.

We have calculated the Engagement Acceptance Rate (EAR), a metric that measures the percentage of fake or artificial engagement (likes, comments, shares, views) accepted by a social network compared to the total ordered engagement. This metric helps conduct a comparative analysis of different platforms based on their ability to prevent engagement manipulation, serving as an important indicator of platform quality.

Analyzing the data on the effectiveness of accepting fake engagement, we found that most types of fake engagement can be easily

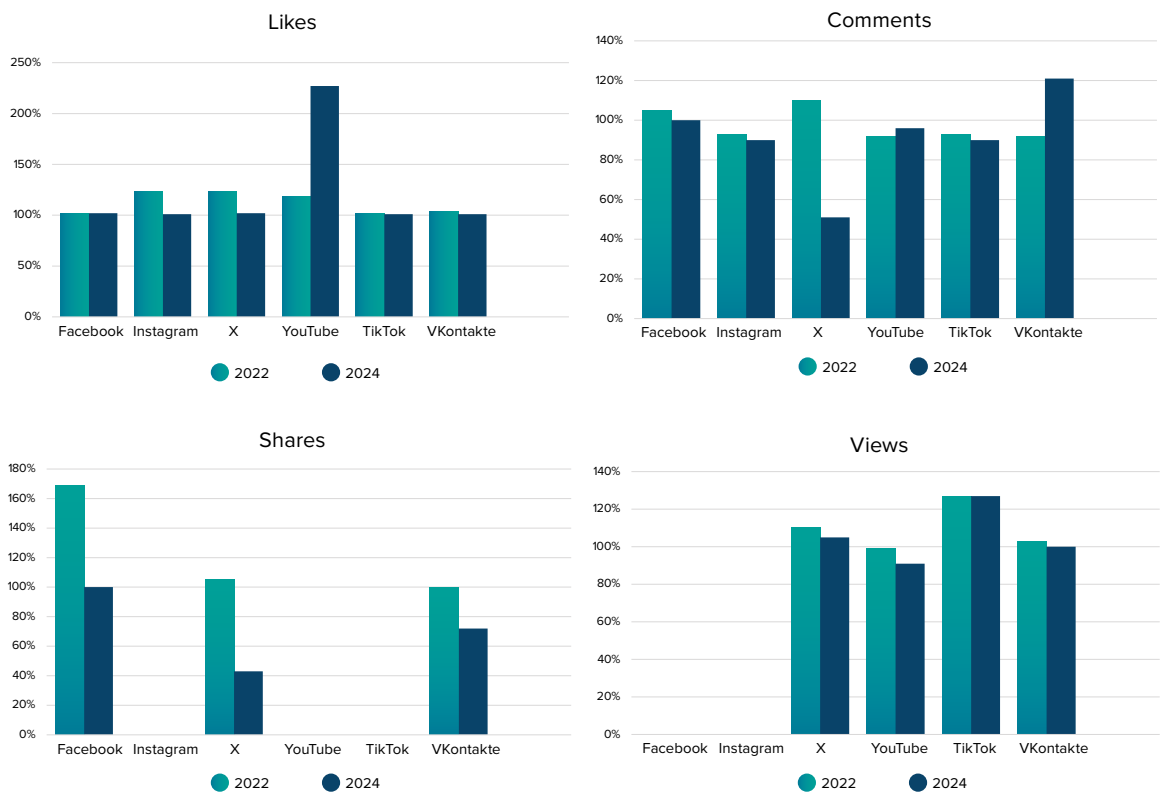


FIGURE 10. Engagement Acceptance Rate (EAR)

manipulated. All platforms allowed for 100% or more fake likes, with YouTube performing particularly poorly in this regard. Social networks also struggled with fake views, where the pass-through rate for this type of engagement ranged from 90% to 115%. Most platforms also underperformed with fake comments, with acceptance rates for this type of engagement ranging from 90% to 121%. However, it should be noted that X handled this type of engagement the best, allowing only 50% of fake comments. The situation is slightly better for shares, where Facebook performed the worst with a 100% pass rate for this type of engagement. Conversely, X again performed the best, permitting only 43% of retweets, while VKontakte allowed 72%. While X has *improved* in comments and shares, it is inefficient in likes, besides X decided to hide what posts were liked by other users, shifting manipulations of those engagements behind the curtain.

So what? We observe poor performance from platforms in terms of countering different types of inauthentic activity. None of the social networks was able to identify and counter fake likes. In the least successful case for manipulation services, more than 90% of fake views were delivered, which represents a significant failure for video-oriented platforms like YouTube and TikTok. The only platform that made significant improvements in combating fake comments and reposts was X, which managed to identify and counter approximately 50% of this activity.

Overview of assessment criteria

In 2024, social media platforms demonstrated varying performance compared to 2022, with some improvements observed in certain assessment criteria for specific platforms. Despite these advancements, platforms continue to struggle in effectively combating commercial bot activity. The overall score, representing the cumulative performance across all criteria (where a **higher score indicates better performance**), shows that Facebook, X, and YouTube have improved, while Instagram has remained stagnant, and VKontakte and TikTok have seen declines. Additionally, social media manipulation services remain readily accessible and have become even more affordable.

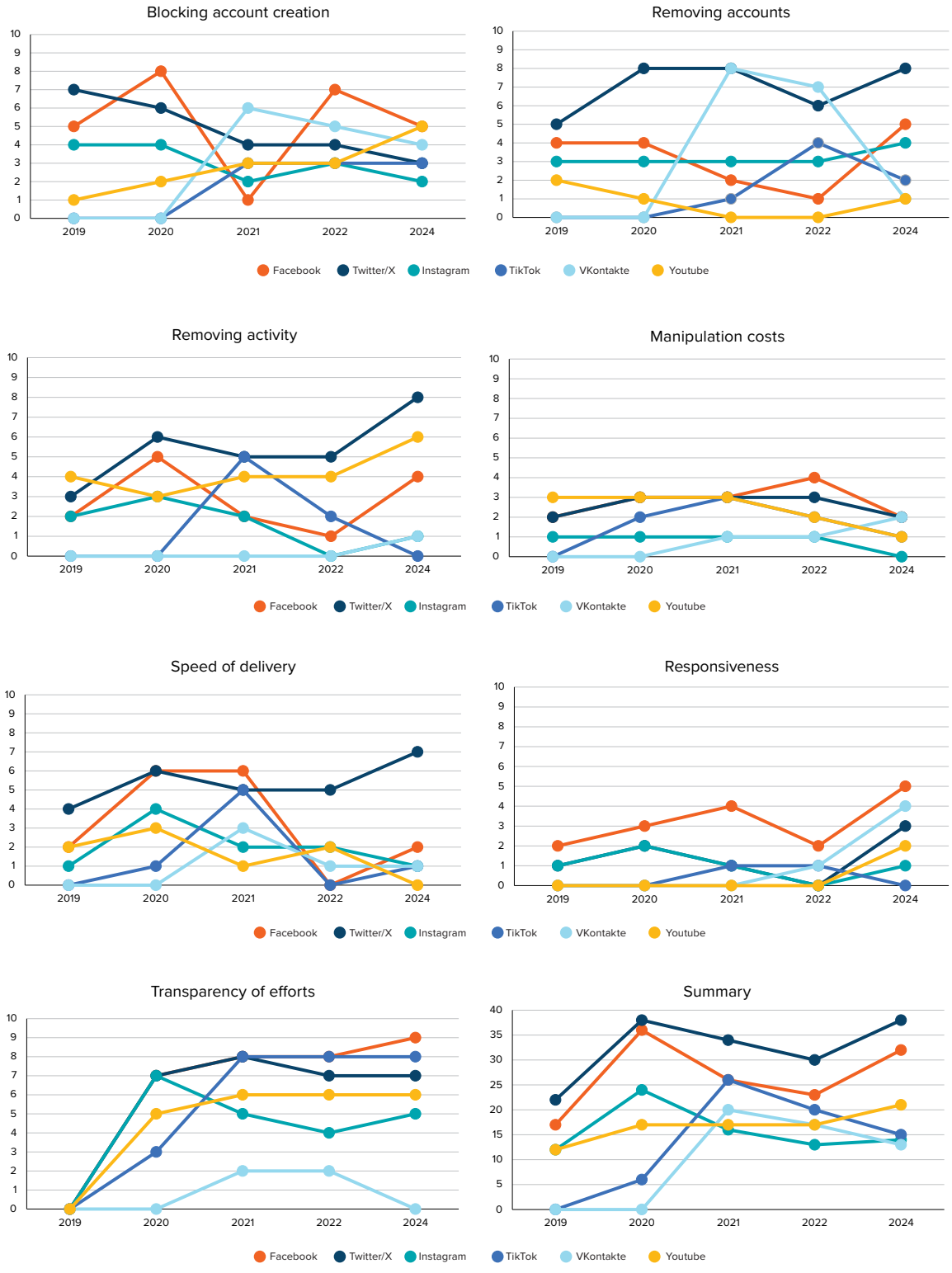


FIGURE 11. Overview of the assessment criteria (scores by platform, 2019-2024)

How does a bot farm operate?

Through our experimentation, we have demonstrated that inauthentic engagement is inexpensive, and social media platforms face challenges in effectively detecting and removing it. But what drives these service providers? To address this question, we will explore how bot farms function and their crucial role in manipulating social media platforms. As all platforms are vulnerable to manipulation through the use of bot networks, we aim to open discussion on threats amplified by automated or semi-automated accounts controlled by "bot farms," accounts that are often deployed to create the illusion of popularity, manipulate public opinion, and disseminate disinformation. While initially used for commercial marketing purposes, the proven effectiveness of these tactics has made them a powerful tool for malicious activities, particularly during elections and political crises. This underscores the broader issue of inadequate regulation and enforcement in the fight against information manipulation.

Bot farms typically undergo several stages in their creation:

Stage 1

The initial stage involves the mass registration of fake accounts, forming the basis for future manipulations in the online space. To circumvent the identification procedures used by social networks or platforms, specialized services are often employed, offering temporary or disposable phone numbers needed for account verification.

Various online platforms are frequently utilized for generating temporary phone numbers to facilitate the creation of accounts on social media platforms such as Facebook, X, Instagram, and even Google services. These services enable users to bypass traditional phone verification processes, making it easier to run bot farms.

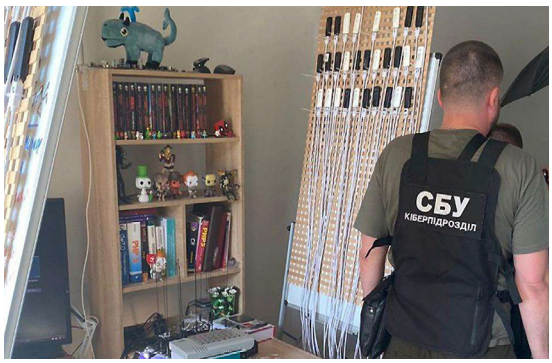
Stage 2

The second stage in the creation process is known as "farming" the accounts. During this phase, newly created accounts are gradually developed to appear more authentic before they are used for manipulations or attacks. The goal is to make these accounts resemble genuine users, preventing their rapid detection and blocking by platforms, which often flag new fake accounts due to suspicious activity or abnormal behaviour. To achieve this, the accounts are populated with activities such as liking posts, adding friends, joining discussions, subscribing to communities, and leaving comments.

Specialised tools, including anti-detect browsers, are employed to facilitate this process. These tools, such as Indigo Browser, Multilogin, GoLogin, and others, allow users to create multiple profiles with different configurations, simulating various devices and browsers, altering User-Agent strings, and concealing real IP addresses and other metadata that could be used to identify bots or fake accounts.

A Notorious Criminal Use of Bot Networks in Ukraine

In 2022, Ukrainian law enforcement dismantled a criminal group led by a Russian citizen that operated a bot farm to discredit Ukraine's leadership and destabilise the socio-political environment during Russia's military invasion. The group spread disinformation online, using bots to circulate fake news about the frontline and conduct influence operations aligned with Russian intelligence interests.



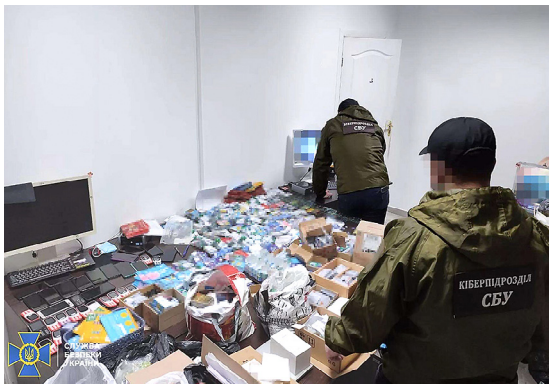


IMAGE 15., 16., 17., 18. Recent photos documenting seized Russian disinformation facilities, showing equipment used for information operations targeting Ukraine's digital space

The Russian national behind the bot farm used custom software to automate the management of hundreds of thousands of accounts. The operation relied on GSM gateways and SIM banks to control large numbers of SIM cards, switching them as needed. Equipment used included multi-port GSM modems, SIM gateways for managing SIM card operations, and SIM banks for storing and handling SIM cards, enabling mass account control and automated messaging.

IMAGE 19. Example of sim-box used for running bot & troll farms

Bots and War

Since the full-scale invasion began, Ukrainian law enforcement and special services have dismantled over 80 bot farms, managing around 10 million fake accounts spreading Russian propaganda. These operations, often coordinated from outside Ukraine, aimed to distribute disinformation via social media. Over 5,500 accounts were blocked, and hundreds of internet agents working for hostile forces were exposed.

The crackdown involved technical analysis to identify shared IP addresses, servers, and suspicious account registration activity, with mobile operators playing a key role by detecting anomalies in network traffic. Financial and domain registration analysis helped identify and dismantle the criminal network behind those bot farms.

This is particularly important for the future battlefield in the information space and highlights the importance of cross-government and cross-institution collaboration. Current information threats are likely to persist globally without significant improvements in the short-term. Understanding how bot farms are used and operated in today's conflict zones will allow the EU, NATO, and their allies to proactively address these challenges. By developing strategies to counter these bot farm operations in the same way as financial fraud operations now, we can enhance our preparedness and resilience against future disinformation and influence campaigns. This will be critical to safeguarding democratic institutions and public trust.

Conclusions & Recommendations

Harmful Bots

Commercial bots might appear relatively harmless due to their focus on topics like cryptocurrency, betting, and similar subjects. However, our observations indicate that these same bot accounts are also utilized to amplify political content, thereby manipulating audience reach mechanisms across various platforms. We observe a very diverse range of topics of interest for puppet-masters behind the generated bot engagements and thus conclude that commercial and spam/scam-driven bots are often reused in political manipulations. With US elections on the horizon, there is little to no doubt that bots will amplify and therefore manipulate election-driven narratives.

Scaling Problem

We concluded that the current red-team approach reveals concerning, stagnant, or even worsening trends across platforms, and there may be various reasons for this. One key reason is the scale of our experiment compared to the immense volume of user activity these platforms handle every day. However, this raises the question: does our experiment still matter? The answer is a resounding yes—more now than ever. Influence is not always driven by large-scale campaigns with millions of engagements. Small, automated, and seemingly smaller-scale disinformation campaigns can be deployed undetected across nearly all platforms. Major social media providers, who are also leaders in AI innovation, should emphasize and prioritize the mitigation of commercial coordination risks as well as language biases in content moderation. In addition to detecting

large-scale operations, it is important to invest in the detection of small-scale multi-lingual coordinated campaigns. These pose a growing threat to the integrity of online spaces when it comes to more efficiently addressing spam/scam and disinformation influence countermeasures in transparency reporting. At this moment we consider this a substantial vulnerability across all platforms. And to conclude this thought, we invite the reader to consider: should transparency reporting be limited to social media platforms, or should it also apply to marketing agencies that conduct campaigns?

Moderation and Collaboration

While AI advancements may enhance the efficiency of content moderation over time, current generative-AI models remain resource-intensive where they typically play part of more complex machine-learning classification systems. But as our experiment indicates, those are far from flawless due to the scaling (vast number of daily user-generated engagement) and possibly other elements such as complexity of languages and data modalities. Therefore, regulation alone cannot solve the problem. Enhanced collaboration with research organizations and institutions provides a complementary approach that prioritizes integrity over profit. By embracing greater transparency and actively engaging with research teams, platforms stand to gain long-term advantages as far as identifying and countering influence operations are concerned. Additionally, regulatory policy frameworks must strengthen the requirement to include mandatory reporting on influence operations and apply stricter penalties for platforms that fail to identify and counter manipulation and coordinated disinformation efforts.

Cheap Manipulation

Manipulation services have significantly decreased in cost, allowing malicious actors to leverage commercial bot activity more efficiently to amplify their campaigns. This suggests that, for the same investment, a much larger scale of inauthentic engagements can now be achieved on these platforms, highlighting the importance of the assessment criteria we have used in the experiment. It is important to point out that different forms of engagement come with varying costs. For instance, comments are significantly more expensive than views or likes, probably because they are more labor-intensive and, in a political context, carry sentiment, opinions, and attitudes that hold more weight. Views, being the cheapest form of engagement, present a particular challenge for TikTok, a video-centric platform, and for researchers, as it becomes difficult to identify bots responsible for generating fake views.

Tailoring Focus

But, are we looking in the right direction? Our five-year study confirms that commercial bot activity primarily serves as an amplification tool in malicious campaigns, rather than constituting their core. Simultaneously, various research studies[5,6,7] highlight the extensive manipulations executed through social media advertisements. Building on this understanding, our next experiment will intensify focus on advertisements, recognising them as a fundamental element of the manipulation ecosystem, alongside commercial bots. We believe it is essential to assess platforms' ability to counter the sources of manipulative content, not just the amplification mechanisms.

Endnotes

1 European Parliament & Council of the European Union. (2022). Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act). Official Journal of the European Union, L 277, 1-102.

2 European Parliament & Council of the European Union. (2022). Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on Contestable and Fair Markets in the Digital Sector and Amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act). Official Journal of the European Union, L 265, 1-66.

3 Cambridge Online Trust and Safety Index *data*

4 To find data on prices for 2023 from the relevant providers, we used a *web archive*.

5 Hernandez, A. (2023, August 1). *Are your ads funding disinformation? Harvard Business Review*.

6 Bergmanis-Korāts, G., Haiduchyk, T., Shevtsov, A. AI in Precision Persuasion. Unveiling Tactics and Risks on Social Media. Riga: NATO Strategic Communications Centre of Excellence

7 Politico. (2024, October 15). *Big, bold and unchecked: Russian influence operation thrives on Facebook*.

Prepared and published by the
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