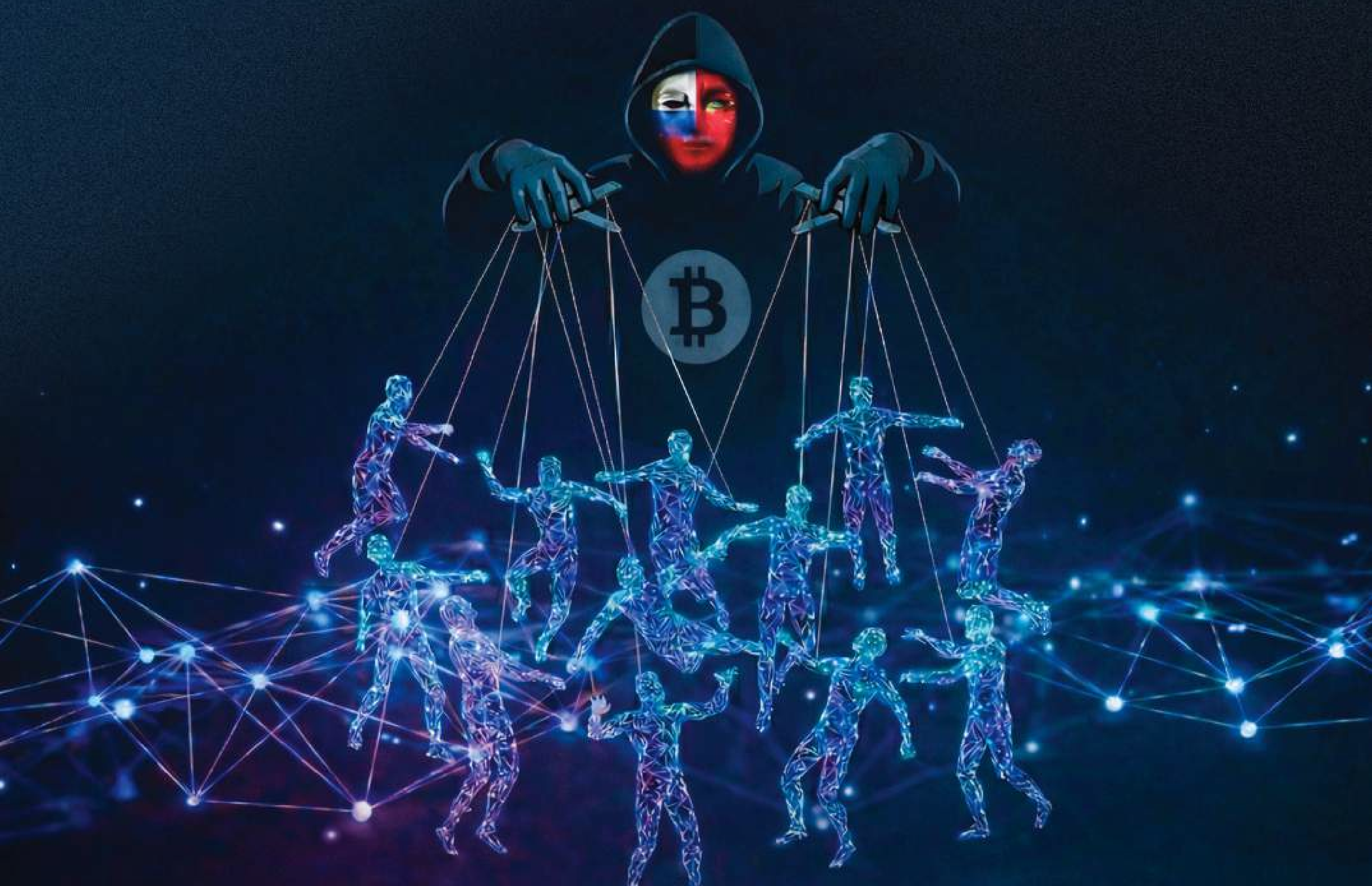
ISBN 978-9934-619-74-8



Social Media Manipulation for Sale

2025 Experiment on Platform Capabilities to Detect and Counter Inauthentic Social Media Engagement

PREPARED AND PUBLISHED BY THE

NATO STRATEGIC COMMUNICATIONS CENTRE OF EXCELLENCE

ISBN: 978-9934-619-74-8

Authors: Dr. Gundars Bergmanis-Korāts, Tetiana Haiduchyk, Bohdan Smolts
Project manager: Dr. Gundars Bergmanis-Korāts
Contributors: Trementum Research, The Financial Intelligence Unit of Latvia, Cyabra
Experiment: Trementum Research
Content Editor: Hadley Newman
Design: Inga Ropša

Cover Image generated by Nano Banana

Riga, January 2026

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

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

Social Media Manipulation for Sale

2025 Experiment on Platform Capabilities to Detect and Counter Inauthentic Social Media Engagement

Contents

Executive summary

This sixth annual evaluation of social media, conducted since 2019, tests the resilience of major social media platforms against manipulation by commercial service providers. The experiment measures platforms' ability to detect and remove inauthentic engagement that is commercially purchased for deliberately created inauthentic posts in non-political scenarios.

Despite the introduction of the European Union's Digital Services Act and Digital Markets Act, and two and a half years after Russia's full-scale invasion of Ukraine, commercial manipulation remains widely available across platforms. The experiment continues to identify persistent vulnerabilities, including the amplification of politically sensitive content. Although X showed notable improvement, removing approximately half of the identified fake accounts and interactions, enforcement outcomes across other platforms remained limited. The persistence of low-cost, easily accessible commercial manipulation services raises concerns about their potential use in amplifying politically sensitive content, even beyond the non-political scenarios tested in this experiment.

This year's findings indicate a mixed picture: enforcement is improving in some areas, but systemic vulnerabilities persist. Several major platforms increased removal activity compared with previous iterations of the experiment, yet commercial manipulation remains inexpensive and easy to obtain. **This year, the experiment expanded to seven platforms and added tests involving sponsored content, AI-generated posts, and a larger set of manipulation providers. Across the tests, more than 30,000 inauthentic accounts delivered[1] over 100,000 units of inauthentic engagement, providing the clearest evidence to date in this series of how manipulation manifests at scale.**

Although platforms removed fake accounts at the highest rate recorded so far, averaging just over half of identified accounts, results varied substantially. Platform performance varied substantially across both account removal and engagement removal. VKontakte and X removed a higher proportion of inauthentic accounts than other platforms, while Instagram and TikTok removed only a small share. A similar pattern was observed for engagement itself: X and YouTube removed the largest proportion of inauthentic activity, whereas Facebook, VKontakte, Instagram, TikTok, and f left the majority of purchased engagement in place.

This year, we expanded the scope of the experiment to include sponsored (advertising) content on Facebook, Instagram, X, TikTok, and YouTube. For a total cost of €252, the campaign generated 206,234 views, 200 likes, and 17,442 inauthentic comments[2]. We also identified a market for ready-to-use inauthentic advertising accounts, successfully purchased for Meta, TikTok, and YouTube. This indicates that commercial manipulation is not limited to organic content and can extend into paid advertising, with the potential for platform ad systems to contribute to the distribution of inauthentic material. Strategically, this shift to paid manipulation allows actors to bypass the skepticism users often apply to organic posts while leveraging platform algorithms to deliver inauthentic narratives to precisely targeted audiences. Although advertisement manipulation appears more expensive and operationally demanding than standard engagement manipulation, it remains feasible at relatively low cost. Observed outcomes varied by platform: Instagram showed the lowest resistance, delivering the highest average volume of inauthentic comments on ad posts (reaching 340% of the expected number[3] after 72 hours). X and YouTube showed partial delivery (approximately 25% and 21%, respectively).

Facebook and TikTok showed stronger resistance in this test, with TikTok showing 0% delivery. Platform transparency and enforcement reporting across major social media platforms remains inconsistent. TikTok stands out in this regard, as it was the only platform to engage directly with the findings of this experiment and to publish detailed information on covert influence operations and enforcement practices during the reporting period. By contrast, other platforms provide limited or outdated transparency reporting, which constrains meaningful cross-platform comparison:

- **X** has not released any transparency or enforcement updates during the period covered by this experiment.

- Meta reports removal actions for Facebook, but does not provide equivalent reporting for Instagram, limiting assessment across Meta-owned platforms.

- **YouTube** and **Bluesky** publish only partial annual figures, restricting visibility into enforcement activity and trends over time.

A significant gap persists between reported enforcement capabilities and routine detection outcomes. While TikTok successfully removed inauthentic activity from posts that were directly escalated to the platform, similar activity was observed to persist on posts that were not escalated. This suggests that reported enforcement capabilities are not yet consistently reflected in routine, at-scale detection. **Ultimately, this uneven transparency obscures the true extent of platform vulnerability and prevents an independent, comprehensive assessment of cross-platform resilience against coordinated manipulation.**

Manipulation services remain easy to obtain and relatively inexpensive, with prices becoming more consistent across platforms. Alongside this emerging behaviour, we observed a notable change in the type of content amplified. While the majority of content amplified by commercial bots still relates to cryptocurrency scams, commercial promotions, and other non-political material, we continue to observe an annual increase in the use of spam bots for promoting political narratives and nation-related issues. Additionally, a **significant increase in military and pro-China themes appeared across several platforms**, with Instagram standing out as the only environment where no such amplification was detected.

According to Cyabra's findings, the behaviour of inauthentic accounts has also emerged in more sophisticated operations. Instead of legacy (such as classical spam bot amplification or commenting) behaviours, where spam bots rely on high-volume spam, **new types of inauthentic accounts are now able to blend into ongoing conversations, using AI-generated text and visuals to appear more convincing and thus appear more authentic.**

Cryptocurrency analysis indicates that **commercial manipulation providers continue to rely on cryptocurrency as their primary payment mechanism due to its speed, cross-border nature, and limited enforceability**. Providers predominantly use custodial wallets and high-risk exchanges, routing customer funds through virtual asset service provider (VASP) hot wallets (e.g., Cryptomus and Heleket) where transactions are commingled, significantly reducing on-chain attribution and making full transaction traceability difficult with only four of ten transactions in the experiment could be reliably traced end-to-end. Despite this low-visibility financial architecture, several operators exhibited substantial transaction volumes, underscoring the scale and persistence of the manipulation market; between November 2023 and November 2025, one Russia-based provider we hereafter referred to as RU1 received an estimated USD 265,261 and another based in UK (referred to as UK2) approximately USD 123,714. **The largely unmonitored nature of this infrastructure reinforces the resilience of the manipulation**

economy and raises potential sanctions compliance concerns, particularly regarding suspected Russia-based operators using major exchange custody under Council Regulation (EU) No. 833/2014, Article 5b(2).

Taken together, the findings indicate that platform defences are improving but remain insufficient. Manipulation remains easy to execute and difficult to reliably prevent. The increasing sophistication of **AI-enabled inauthentic accounts allows them to influence conversations with a lower likelihood of detection, particularly when activity is divided into small, distributed actions**. The financial infrastructure supporting these services also remains largely unmonitored, reinforcing the resilience of the manipulation economy.

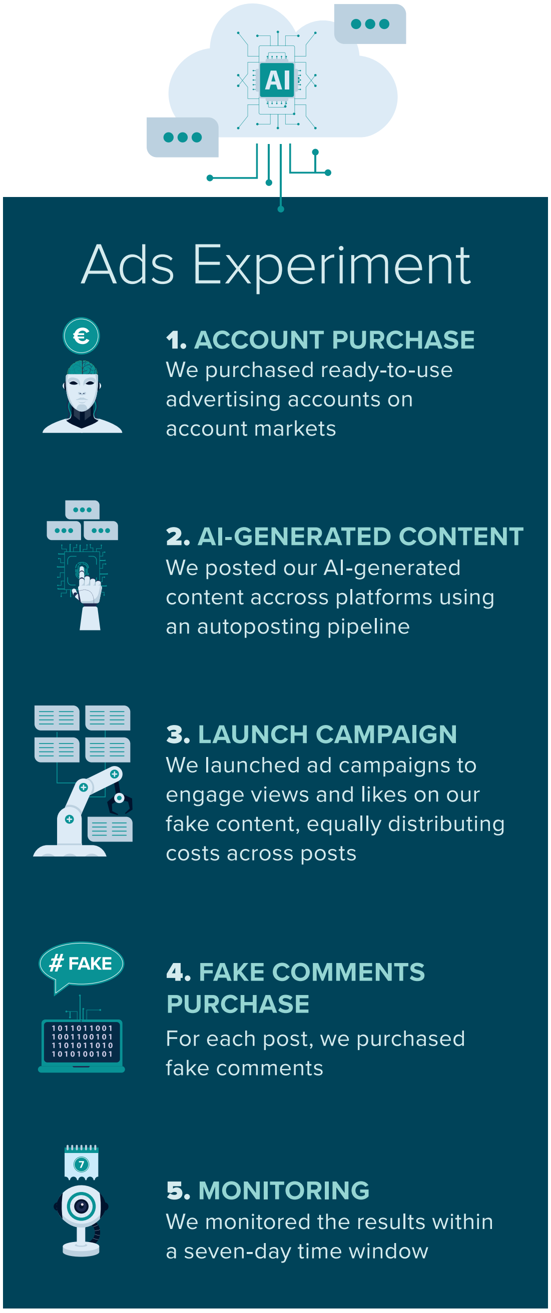
The findings of this experiment suggest that effectively countering these trends requires a shift towards behavioural detection methods focused on timing patterns, account relationships, and coordinated activity across platforms and environments. Analysis indicates that enforcement is more effective when shifting from isolated campaign responses to continuous, context-driven monitoring. Furthermore, the data highlight the importance of analysing entire conversations rather than individual posts to identify bots operating within genuine discussions. Finally, the results underscore that stronger cooperation with financial intelligence units represents a strategic pathway to identifying manipulation providers and limiting their operational capacity.

The experiment

Background

Between September and November 2025, we conducted an experiment to test the ability of social media platforms to identify and remove manipulation. This annual red-team experiment tests and assess social media platform resilience against manipulation. Specifically, it focuses on manipulation orchestrated by commercial service providers. A key restriction this year remains unchanged from previous reports, as we limit the use of commercial manipulation services (purchased inauthentic engagement) exclusively 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 (followers, likes, views, shares and comments) on 126 deliberately created inauthentic posts we have created using 28 inauthentic 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 a total expenditure of €252, we purchased predefined engagement "baskets" offered by commercial manipulation providers with standardised service packages that typically include fixed quantities of likes, comments, views, and followers (e.g., 100 likes, 100 comments, 1,000 views, or 100 followers or friends). Prices varied by platform and provider, and purchases were made across multiple platforms, resulting in 17,553 inauthentic comments, 37,814 likes, 16,025 shares, and 27,653 views delivered on Facebook, Instagram, YouTube, TikTok, VKontakte, X, and Bluesky. In total, we identified 30,011 unique inauthentic social media accounts responsible for generating this engagement, with the expenditure reflecting the cumulative cost of these predefined packages rather than a target-driven engagement volume.



Ads Experiment

1. ACCOUNT PURCHASE
We purchased ready-to-use advertising accounts on account markets

2. AI-GENERATED CONTENT
We posted our AI-generated content accross platforms using an autoposting pipeline

3. LAUNCH CAMPAIGN
We launched ad campaigns to engage views and likes on our fake content, equally distributing costs across posts

4. FAKE COMMENTS PURCHASE
For each post, we purchased fake comments

5. MONITORING
We monitored the results within a seven-day time window

Additionally, we conducted an Ads experiment on Facebook, Instagram, X, TikTok, and YouTube. For €130, we received 206,234 views and 200 likes from ad campaigns launched with purchased fake accounts. Then, for €121, we received 17,442 comments on the same posts through our ads experiment. Rationale for ads: Given that various investigations, including investigations by Reuters[4], have identified inauthentic advertisements as a significant issue for the social media platforms, this report assesses the ease with which these ads can be deployed and manipulated.

Improvements

- **Platforms:** The 2025 methodology expands the platform set by adding **Bluesky** to Facebook, X, Instagram, TikTok, YouTube, and VKontakte.
- **Scale:** We increased the number of **inauthentic accounts** created and **the** volume of **purchased manipulation**. We also expanded the pool of **manipulation providers to ten**.
- **Content type:** In addition to regular posts, we expanded the 2025 experiment to include **sponsored content**.
- **Content origin:** Alongside manually created content, we incorporated multiple types of **AI-generated content** and **automated cross-platform posting** for one of our AI personas.

Assessment criteria

This report analyses social media platforms against key criteria related to countering inauthentic activity. The assessment focuses on platform effectiveness in preventing the creation of fake accounts and their capacity to detect and remove both these accounts and associated manipulative actions. A crucial part of the evaluation is the accessibility and cost of manipulation services across various regions. Additionally, we evaluate the speed and extent of manipulation, the platforms' reaction time to these threats, and the clarity and openness of their actions and reporting. These criteria are operationalised in the following sections through specific metrics used to compare the capabilities of social media platforms to identify and counter commercial manipulation, including successful account creation rates, detection latency (time-to-removal), and a comparative analysis of black-market service pricing.

Blocking the creation of inauthentic accounts

We assessed platforms' ability to prevent the creation of inauthentic accounts from several angles: account creation process, the ability to purchase ready-to-use advertising accounts, the market prices of inauthentic accounts, and the SMS verification services. Researchers describe SMS verification infrastructure as a cornerstone of the online manipulation economy[5]. Analysing this infrastructure helps to assess the effectiveness of social media platforms in preventing the creation of inauthentic accounts.

The first layer of assessment focused on the account creation process itself, using accounts that we later publish posts from for purchased inauthentic engagement. This year, we created four accounts per platform (two in previous years).

In 2025, the procedure for creating inauthentic accounts remained unchanged from previous experiments. However, Facebook, YouTube, and TikTok responded to the increase in accounts in different ways. Facebook fully blocked newly created accounts, YouTube requested additional phone verification, while TikTok temporarily restricted mobile-based account registration. In contrast, we encountered no challenges when creating inauthentic accounts on Bluesky.

At this stage, we also tested platforms' ability to detect and counter automatically created accounts. Using SMS and email verification service APIs, we built a functioning pipeline for automated account creation on TikTok, Instagram, X, and YouTube. TikTok and YouTube showed relatively low resistance to automated account creation because registration could be completed without CAPTCHA (a challenge-response test used to distinguish humans from automated systems). In contrast, X and Instagram showed higher resistance: while account creation could still be automated, it typically required additional steps-most notably CAPTCHA-often handled via a third-party solving service. In this comparison, our primary threshold for "resistance" is whether an account can be registered fully automatically. Among platforms where automated registration remained possible, we further compared relative difficulty (i.e., additional effort/complexity required). By this secondary criterion, X and Instagram imposed more friction than TikTok and YouTube.

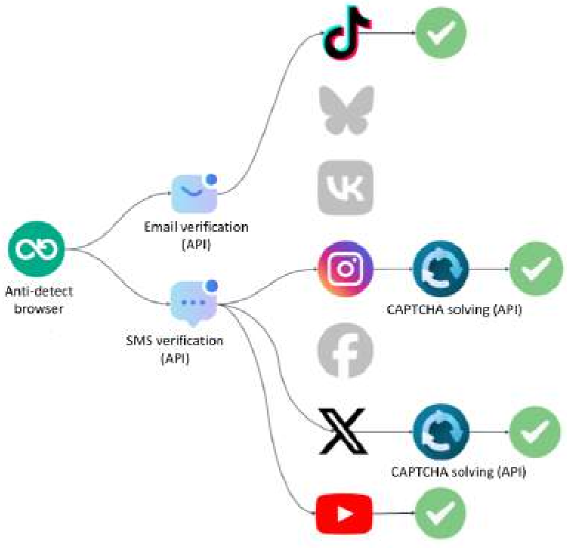


FIGURE 1. Accounts autocreation pipeline

In contrast, Facebook and Bluesky detected and suspended our accounts immediately after creation. We were unable to automate account creation on VKontakte within this test, as the platform allows registration only through the mobile application, which requires more complex automation involving AI-enabled phone automation.

An additional layer of assessment focused on the ability to purchase ready-to-use advertising accounts on social media platforms. The availability of such a market indicates that manipulation is not limited to organic content but extends to paid advertising. Unlike regular content, where spam bots can serve as amplifiers, in the advertising context, the platform itself becomes the amplifier by actively distributing the content. We successfully purchased ready-to-use inauthentic advertising accounts for Meta (Facebook and Instagram), TikTok, and YouTube. For X, we purchased a standard inauthentic account and manually configured the Ads Manager. The experiment excluded Bluesky and VKontakte from this segment. Bluesky was excluded because it lacks an ads manager, and VKontakte requires complex government-curated identity verification procedures.

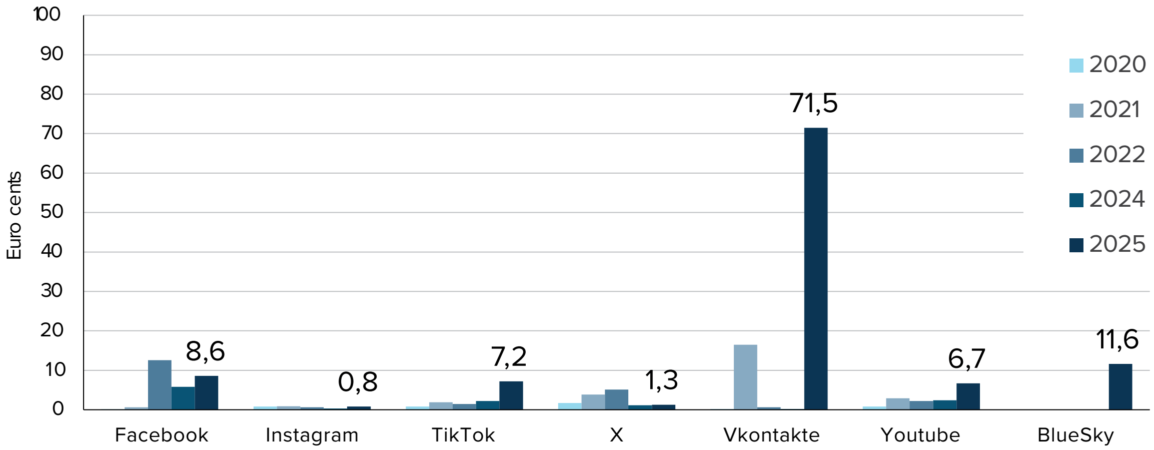


FIGURE 2. Cost of inauthentic accounts in Euro cents grouped by platforms

Further, our assessment examined the market prices of inauthentic accounts and SMS verification services, which are commonly used to create such accounts. Our findings suggest that the higher prices are not a reflection of demand but also an indication of stronger efforts at the platform level to prevent the creation of inauthentic accounts. We analysed average account prices across three markets. Compared with 2024, VKontakte showed the most significant price increase, followed by YouTube and TikTok. Facebook saw a moderate rise, while Instagram and X remained the cheapest. Overall, prices increased across all platforms, although purchasing fake accounts remains relatively cheap.

For the SMS verification prices comparison, we analysed data from the Cambridge Online Trust &Safety Index[6] which includes data from over 500 manipulation services. Compared with 2024, minimum SMS verification prices increased for YouTube, Facebook, and Instagram, while they decreased for VKontakte and X. TikTok remained unchanged. YouTube remains the most expensive overall.

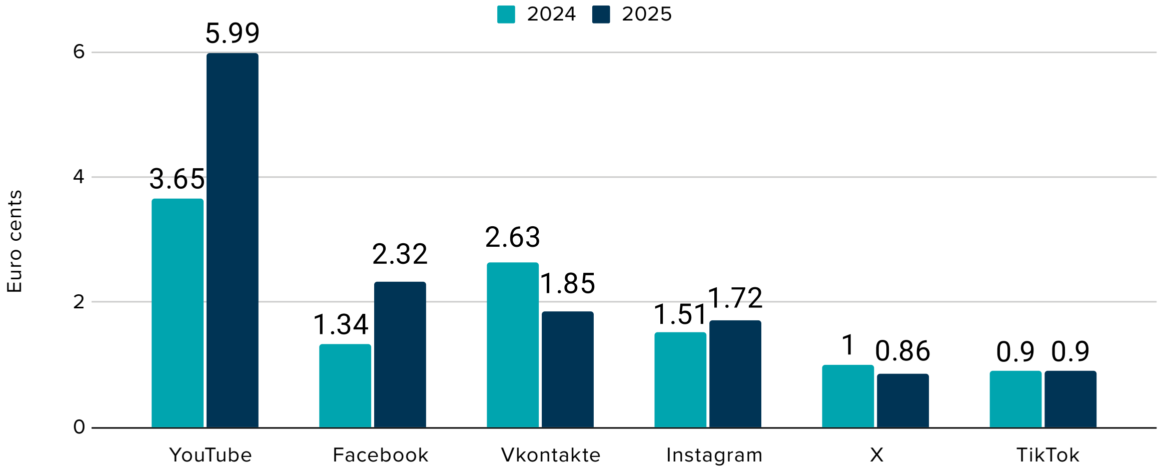


FIGURE 3. Minimal SMS verification price across platforms (2024 vs 2025)

Our advertising experiment revealed a significant price disparity between inauthentic and advertising-ready accounts, with YouTube accounts showing the largest difference.

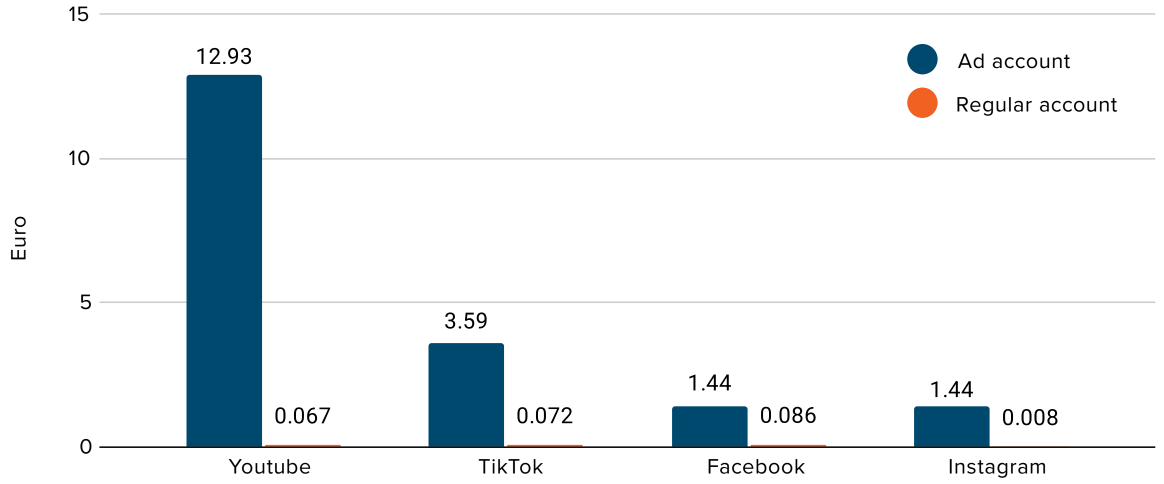


FIGURE 4. Prices of inauthentic accounts configured for ad launching versus regular accounts. We excluded X from this comparison because readily available ad accounts for this platform could not be found and purchased.

Furthermore, YouTube required the longest time for ad approval among all platforms tested. While this indicates that ad manipulation is more difficult compared to regular manipulation and is also more expensive, overall, it remains relatively inexpensive and feasible.

Removing inauthentic accounts

We checked the availability of all identified accounts responsible for delivering fake engagement five weeks after the purchase.

This year, an average of 50.4% of identified inauthentic accounts were removed, representing the highest removal rate observed in this experiment. By comparison, average removal rates in previous rounds remained substantially lower, at approximately 20% in 2021, 14% in 2022, and 13% in 2024, underscoring a marked increase in enforcement activity this year.

This year, VKontakte showed the highest removal results, removing 96% of inauthentic accounts. However, fake engagement from deleted bots remained active. X also performed strongly with 82% removed. YouTube and Bluesky each removed 55% of fake accounts, showing a substantial improvement compared to previous years. Facebook removed 39%, while Instagram and TikTok had lower removal rates of 22% and 4%, respectively.

Several of our accounts created for posting inauthentic content were taken down by the platforms two months after the launch of our experiment. Bluesky removed all accounts, and X removed two out of four that had engaged in fake follower activity. In both cases, platforms stated that the reason for removal was inauthentic behaviour.

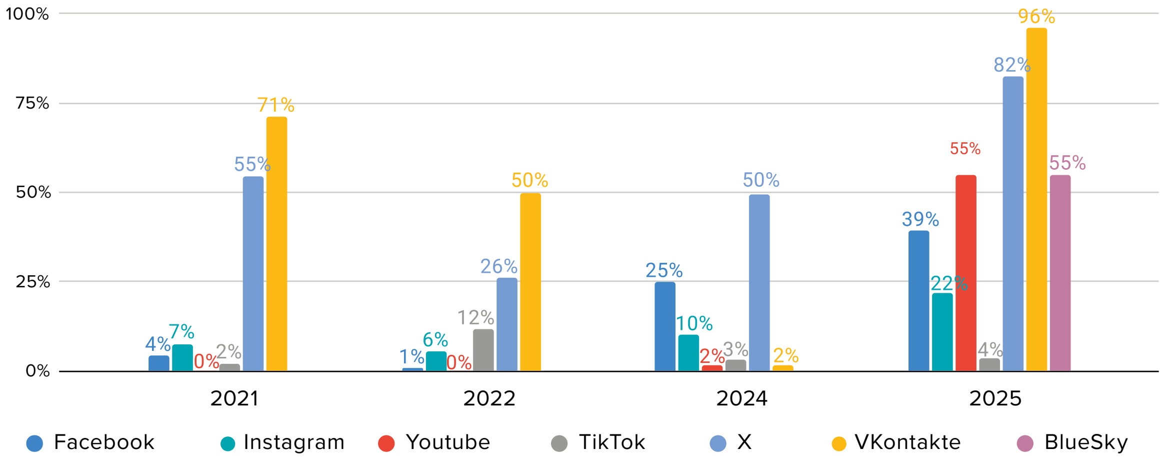


FIGURE 5. Removed inauthentic accounts during the monitoring period

Removing inauthentic activity

> The volume of fake engagement was monitored for each post over four weeks after the purchase.

In 2025, X and YouTube showed the strongest performance. After four weeks, only 43.41% of fake engagement on X and 56.20% on YouTube was still online. Facebook and VKontakte also performed better than in 2024, with 79.40% and 70.35% of fake engagement remaining. Instagram and TikTok showed weaker results, with 84.03% and 83.13% of inauthentic activity still active. BlueSky showed no reduction at all with 100% of fake engagement remained. Overall, most platforms improved their removal of fake engagement compared to previous years.

	2020	2021	2022	2024	2025
Facebook	96.53%	98.52%	99.49%	93.48%	79.40%
X	74.23%	83.43%	82.27%	61.91%	43.41%
Instagram	91.80%	96.01%	99.94%	98.62%	84.03%
TikTok	99.69%	84.77%	97.33%	99.85%	68.13%
VKontakte		99.96%	99.92%	99.32%	70.35%
YouTube	97.17%	92.38%	90.00%	78.71%	56.20%
Bluesky					100.00%

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

Cost of services

The data for Bluesky was collected separately, as this platform is not available among the providers used for cross-platform price comparison in the experiment.

In 2025, prices for social media manipulation varied across platforms. X became the most expensive, showing a clear increase from the previous year. YouTube prices also increased slightly, while Facebook prices continued to drop, following the trend seen in recent years. TikTok and VKontakte prices decreased. Instagram showed a small increase but remained the third-cheapest platform. Bluesky was the most affordable overall.

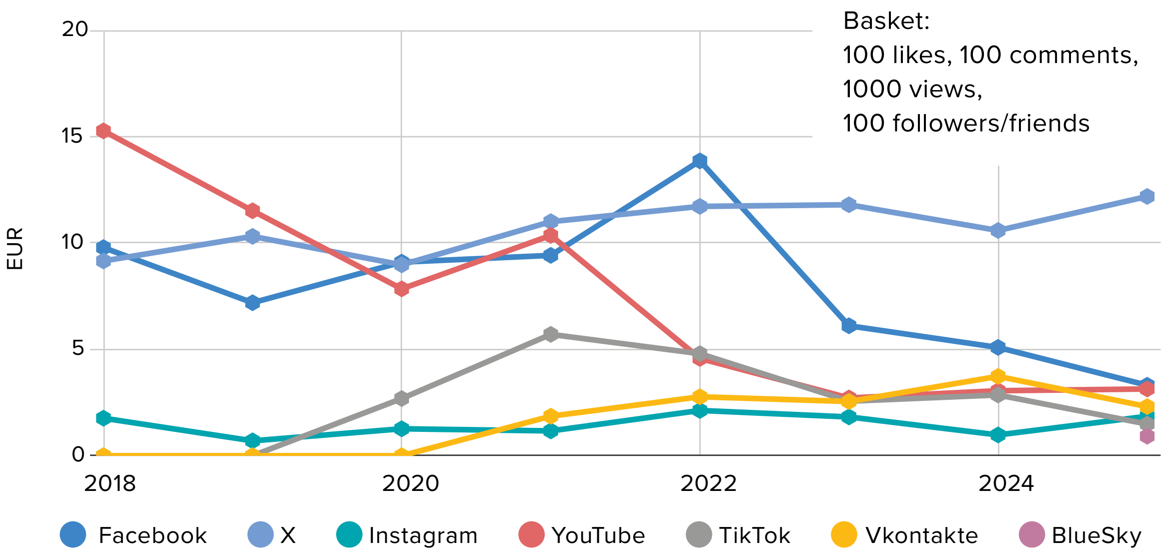


FIGURE 6. Price of a basket of social media manipulation

In 2025, the overall cost of manipulation across platforms generally converged. The notable exception was X, which saw an overall price increase for manipulation except for views which dropped compared to the previous year, making it possible to acquire 156,000 fake views for just €10. This price point keeps X views comparable to and affordable alongside other types of engagements available on other social media platforms.

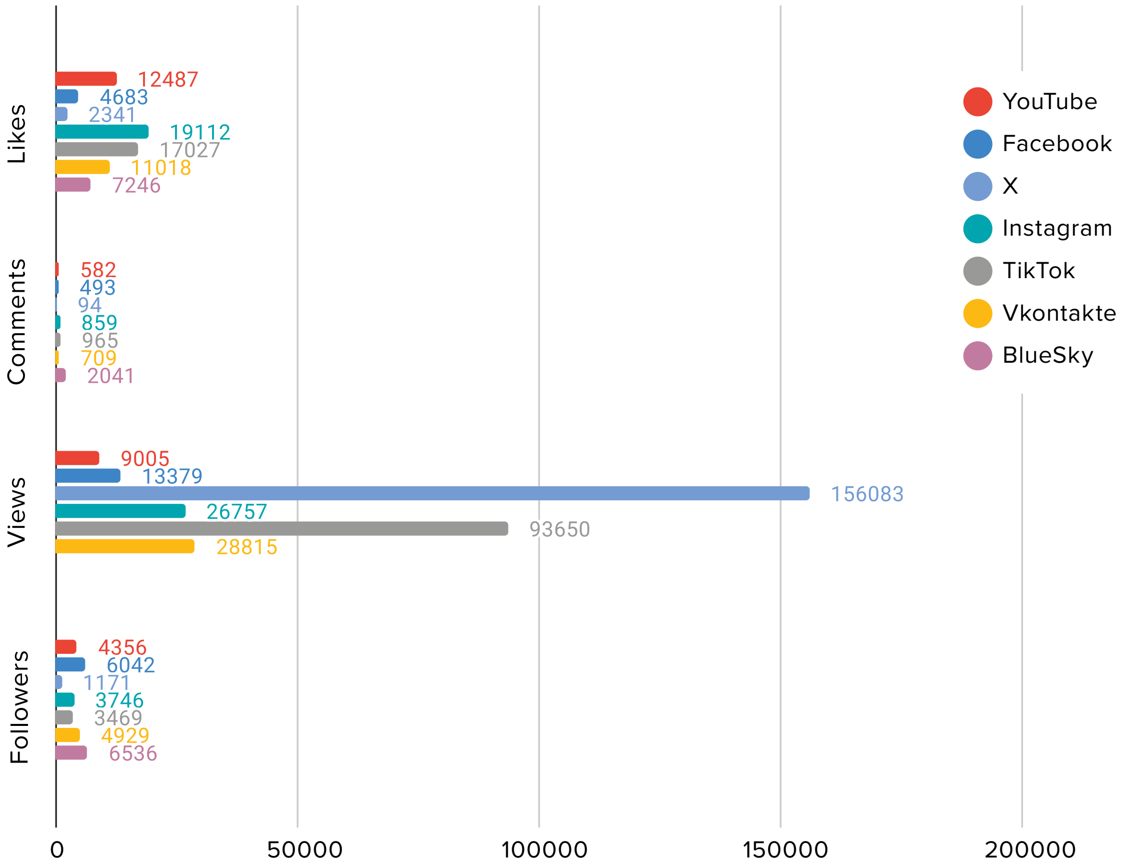


FIGURE 7. How much manipulation can you buy for 10 EUR?

Paid advertising, despite its higher cost compared to regular manipulation, demonstrated a significant potential for reach even with a small budget (e.g., 10 EUR). However, engagement remained disproportionately low relative to the expense.

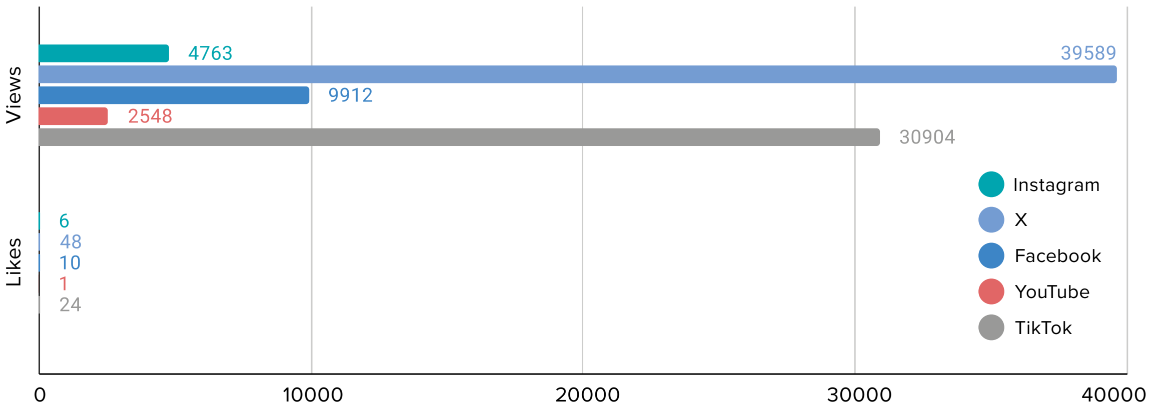


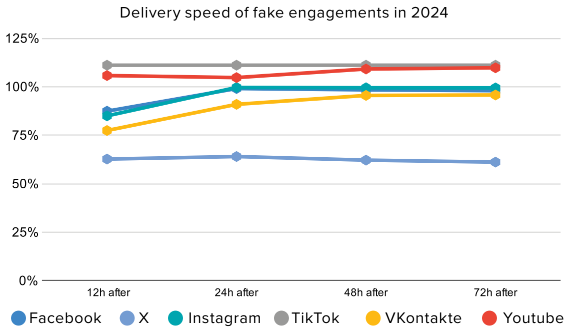
FIGURE 8. How much manipulation can you buy with 10 EUR using ads? The chart illustrates the engagement generated during the experiment through purchased inauthentic advertising accounts.

Volume of engagements initially delivered (speed)

We measure volume vs. the speed in the initial phase of the experiment by assessing the share of ordered fake engagement that was delivered within the first 72 hours after purchase. Slower delivery is interpreted as a sign of better platform detection and removal of inauthentic activity, as it was observed during the testing of various commercial manipulation providers.

In 2025, fake engagement was delivered more slowly than in the previous year. While in 2024, on average, about 96% of all manipulative activity was completed within the first 72 hours, this year the average dropped to around 76.5%.

2024



2025

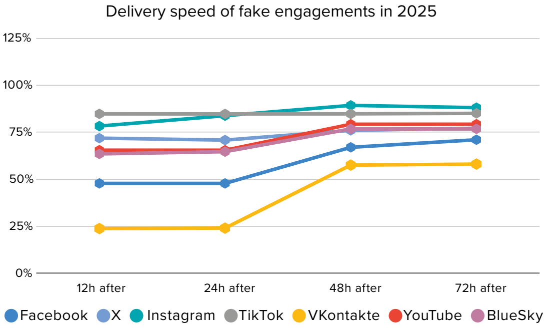


FIGURE 9. Unlike last year, VKontakte showed a significant change. In 2025, it became the platform where fake engagement took the longest to be delivered.

Platforms such as **TikTok and Instagram allow the delivery of fake engagement the fastest**, with most of their activity occurring within 2 days. X and YouTube show moderate delivery speeds, while Facebook remains slower, and VKontakte continues to lag behind all other platforms. Overall, fake engagement in 2025 appears later compared to 2024.

AI-generated vs. Manually Created Content: Assessing the Impact of Engagement

To compare platforms' ability to identify and counter fake activity between AI-generated and manually created content, we distributed content equally between our inauthentic accounts and purchased the same amount of fake engagement for each.

The effectiveness of AI-generated content compared to manually created content varies across different platforms, suggesting inconsistent results following initial testing. In fact, AI-generated content tends to receive fake engagement more quickly on X and Instagram, while for manually created content engagement was delivered faster on Facebook, YouTube, and VKontakte.

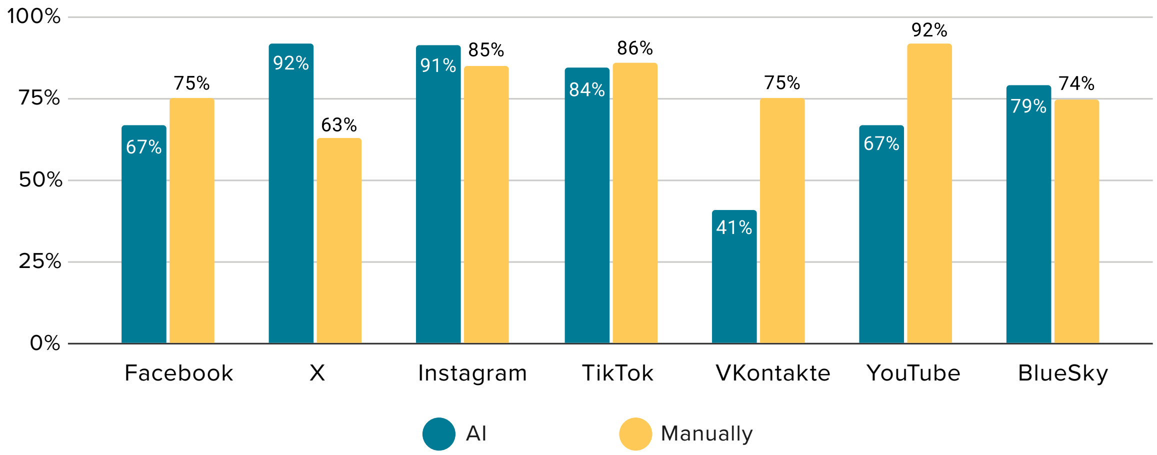


FIGURE 10. Delivery speed of fake engagement for AI-generated vs manually Created content within 72 hour window

The findings suggest that particularly **these three platforms are relatively more effective at identifying and taking down or at least decreasing the scale of automated AI-generated manipulations.**

Manipulation of adverts

For the ads experiment, we only purchased fake comments, as likes and views were expected to come naturally from the ads we launched on AI-generated posts. Overall, after 72 hours, Instagram showed the highest average delivery of fake comments, YouTube and X had moderate levels, while Facebook and TikTok demonstrated strong resistance to fake comments on ad posts.

After three days, Instagram had the highest activity, **reaching an average of 340%**, meaning more than three times the expected (purchased) number of fake comments were delivered. YouTube followed with an average of 21%, showing limited but noticeable delivery. X reached around 25% on average. Facebook remained almost inactive, with less than 1% of fake comments delivered. TikTok showed 0%, meaning no fake comments appeared at all.

Responsiveness to reporting

Five weeks after the purchase, we reported a sample of the fake accounts responsible for delivering the purchased engagement.

After reporting a sample of inauthentic posts, our findings suggest that, when it comes to removing fake accounts, platforms rely more on their internal detection systems than on user reports. Overall, none of the platforms managed to remove more than a quarter of the reported fake accounts.

In 2025, the removal rate of reported accounts increased notably compared to previous years. Facebook showed the greatest improvement, removing nearly a quarter of the reported profiles.

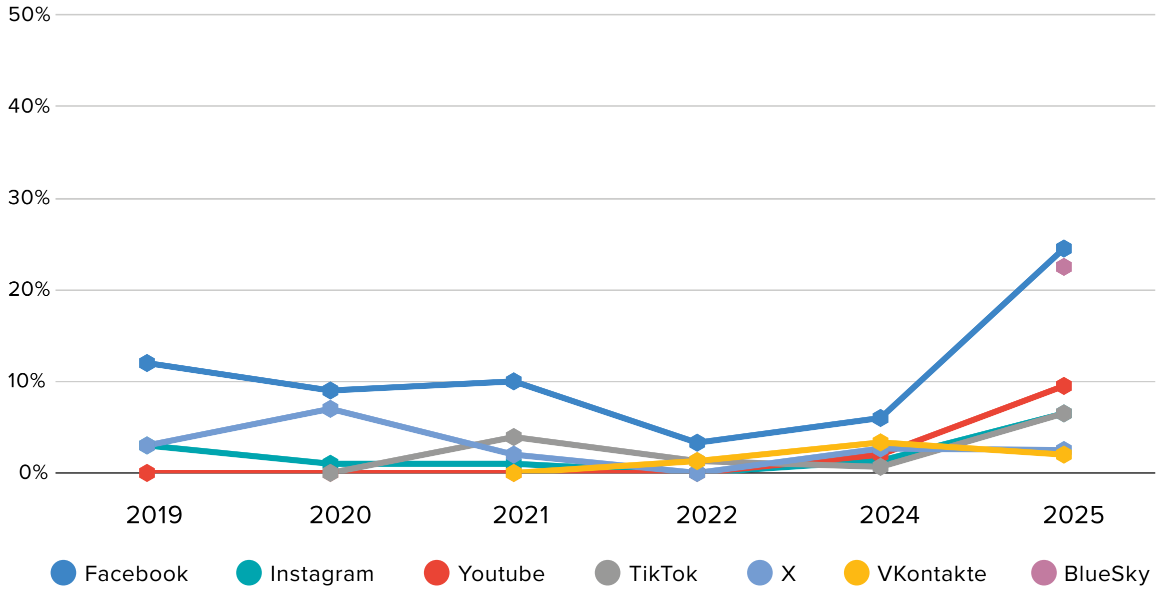


FIGURE 11. Share of accounts removed five days after reporting (21 days after for 2019 assessment).

BlueSky also demonstrated relatively good progress in account moderation, ranking second among all platforms. YouTube, TikTok, and Instagram demonstrated moderate increases, while VKontakte and X showed the least effectiveness, removing only 2% and 2.5% of accounts, respectively, which was even less than for last year's experiment.

Transparency of actions

To evaluate the transparency of platforms' enforcement practices, we examined their public reports on actions taken against inauthentic behaviour. In addition, at the end of the experiment, we contacted the platforms to obtain their perspectives on the findings.

Platforms continue to apply different approaches to reporting on inauthentic behaviour. ***TikTok*** remains the leader, offering the most transparent reports that include not only the number of accounts removed for violating Community Guidelines, but also detailed disclosures on Covert Influence Operations. Significantly, **TikTok was the only platform to respond to our transparency request**. They agreed to a discussion and provided clarification on their performance.

X's latest available report covers activity from 2024, with no updates published this year. ***Meta*** continues to report only the number of removed accounts on Facebook, but does not provide equivalent reporting for Instagram.

Meanwhile, ***YouTube*** keeps reporting the number of channels removed for spam, misleading content, or scams. ***Bluesky*** also reports annually on the number of accounts removed for reasons such as spam and bot networks.

According to TikTok, it aggressively combats inauthentic accounts and fake engagement through a comprehensive strategy that includes platform policies, dedicated teams, machine-learning models, and monitoring of the resale market on a massive scale. The company highlights measurable year-over-year improvement in its enforcement, with internal tracking confirming this progress. As evidence of this improvement, TikTok refers to this year's assessment, which reportedly showed a significantly higher fake-engagement removal rate compared to the findings in our previous report. Finally, TikTok contends that manipulation trends are driven by broader market forces and evolving attacker methods. Therefore, its priority is maintaining real-time defences and unified detection across organic content, advertisements, and AI-generated media, in addition to enforcing mandatory AI labeling and restrictions on harmful content. Later on, a sample of manipulated posts shared with TikTok resulted in the removal of the inauthentic activity. However, the inauthentic activity in the remaining posts, which were not directly shared with TikTok, was not addressed.

Overview of assessment criteria

In 2025, social media platforms demonstrate performance improvement compared to 2024. However, this year we also conclude that despite these advancements, platforms continue to struggle in effectively combating commercial bot activity.

YouTube, TikTok, and VKontakte showed the most noticeable performance gains, contributing the most to the overall score increase (indicating better performance) compared to the last assessment period. Following closely, Instagram and Facebook also improved, though to a slightly smaller degree. Conversely, the overall score for X saw a slight decline this year. This drop is attributed to insufficient transparency, a lack of responsiveness, and no progress in addressing account blocking issues. Additionally, services enabling social media manipulation remain easily accessible and affordable.

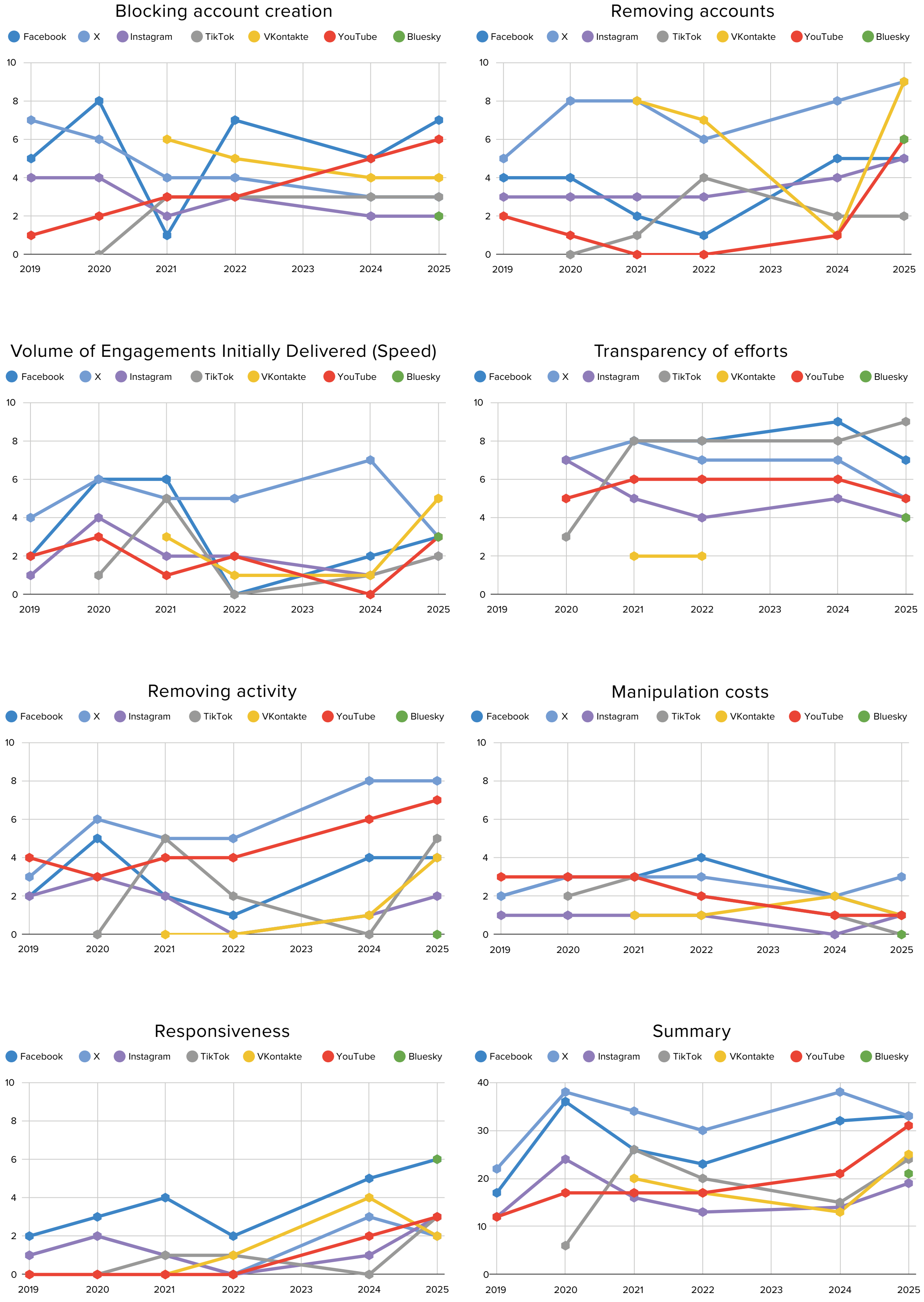


FIGURE 12. Overview of the assessment criteria (scores by platform, 2019-2025)

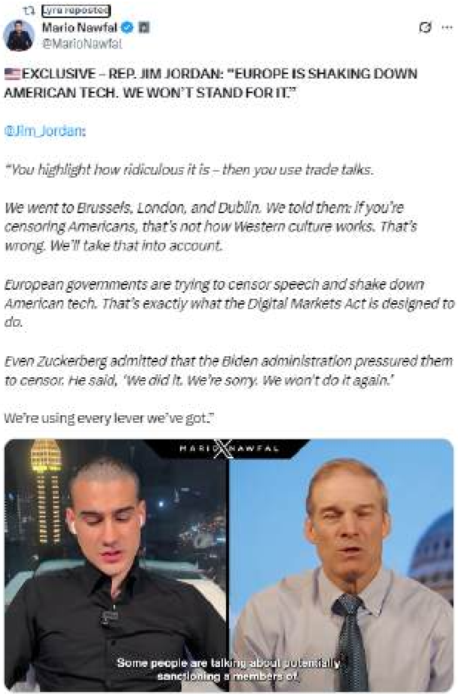
A significant and unexpected development this year was the success of VKontakte in eliminating a substantial number of engagements and accounts. **This performance is noteworthy and could be viewed as consistent with broader efforts to strengthen digital control, although the underlying drivers cannot be determined from this experiment alone.** The long-term effect of this trend on the platform's capacity to moderate its content and overall user engagement capabilities is still unclear.

Recycled spam-bots as amplifiers

This section presents the activity of the accounts that delivered fake engagement in our experiment. As highlighted in previous reports, the bots identified in this context act as amplifiers of manipulation. Therefore, by examining the other content these accounts helped to distribute, we gain insights into the actors behind the use of manipulation services.

In 2024, we observed a growing diversity of political content promoted by commercial bot accounts. This trend persisted in 2025. On X, the spam bots we identified reposted content critical of the European Digital Services Act, as well as posts critical of the *Biden administration and the Democratic Party*.

On Bluesky, in turn, we identified several accounts followed by bots from our sample that shared content critical of Donald Trump and his MAGA campaign.



Lyra reposted

Mario Nawfal
@MarioNawfal

🇺🇸EXCLUSIVE – REP. JIM JORDAN: "EUROPE IS SHAKING DOWN AMERICAN TECH. WE WON'T STAND FOR IT."

@Jim_Jordan:

"You highlight how ridiculous it is – then you use trade talks.

We went to Brussels, London, and Dublin. We told them: If you're censoring Americans, that's not how Western culture works. That's wrong. We'll take that into account.

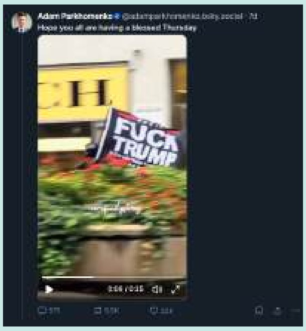
European governments are trying to censor speech and shake down American tech. That's exactly what the Digital Markets Act is designed to do.

Even Zuckerberg admitted that the Biden administration pressured them to censor. He said, 'We did it. We're sorry. We won't do it again.'

We're using every lever we've got."

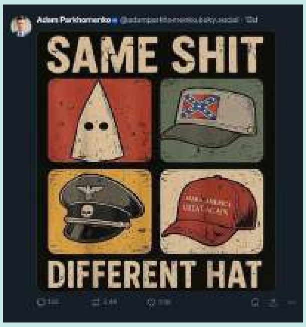
MARIO NAWFAL
Some people are talking about potentially sanctioning a members of.

https://x.com/MarioNawfal/status/1964066003673108690



Adam Parkhomenko @adamparkhomenko.bsky.social · 7d
Hope you all are having a blessed Thursday

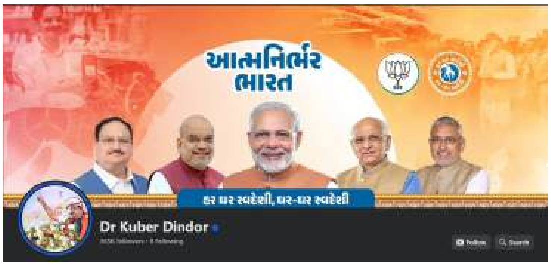
https://bsky.app/profile/adamparkhomenko.bsky.social/post/3m3vmz2xxjk2m



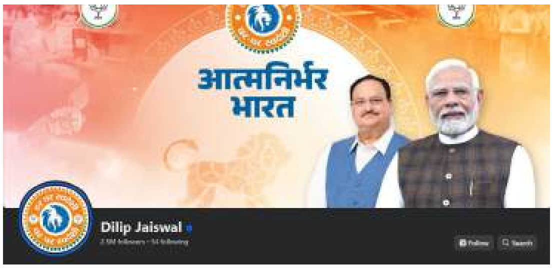
Adam Parkhomenko @adamparkhomenko.bsky.social · 13d

https://bsky.app/profile/adamparkhomenko.bsky.social/post/3m3gr6ikrnk2r

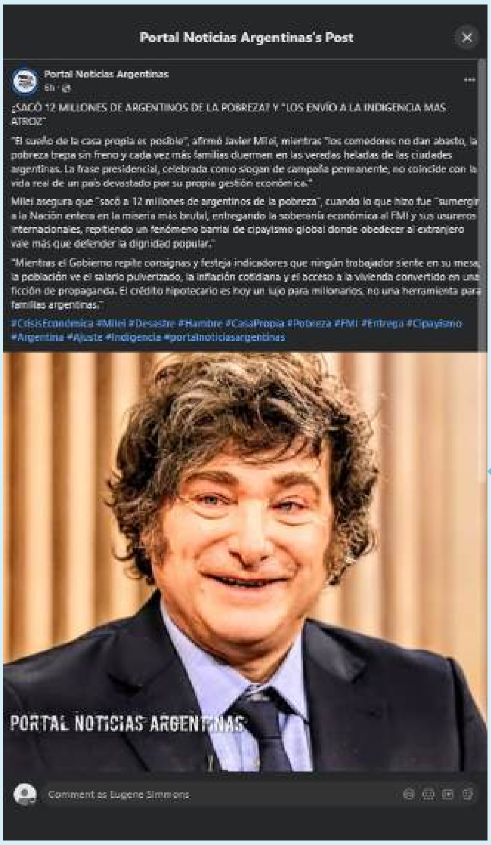
The spam bots in our sample displayed engagement with a range of political subjects, as evidenced by their strong following of both Indian politicians' accounts on Facebook and pages critical of the Argentine government and President Javier Milei.



https://www.facebook.com/kuberdindor



https://www.facebook.com/DilipJaiswalBJP



Portal Noticias Argentinas
6h ·

¿SACÓ 12 MILLONES DE ARGENTINOS DE LA POBREZA? Y "LOS ENVIÓ A LA INDIGENCIA MÁS ATROZ"

"El sueño de la casa propia es posible", afirmó Javier Milei, mientras "los comedores no dan abasto, la pobreza trepa sin freno y cada vez más familias duermen en las veredas heladas de las ciudades argentinas. La frase presidencial, celebrada como slogan de campaña permanente, no coincide con la vida real de un país devastado por su propia gestión económica."

Milei asegura que "sacó a 12 millones de argentinos de la pobreza", cuando lo que hizo fue "sumergir a la Nación entera en la miseria más brutal, entregando la soberanía económica al FMI y sus usureros internacionales, repitiendo un fenómeno barrial de cipayismo global donde obedecer al extranjero vale más que defender la dignidad popular."

"Mientras el Gobierno repite consignas y festeja indicadores que ningún trabajador siente en su mesa, la población ve el salario pulverizado, la inflación cotidiana y el acceso a la vivienda convertido en una ficción de propaganda. El crédito hipotecario es hoy un lujo para millonarios, no una herramienta para familias argentinas."

#CrisisEconómica #Milei #Desastre #Hambre #CasaPropia #Pobreza #FMI #Entrega #Cipayismo #Argentina #Ajuste #Indigencia #portalnoticiasargentinas

https://www.facebook.com/InfoMerlook/posts/pfbid02gPgYnjC4Z81Jqk-d7yKeVm9ga42PE2qjgLdLE1yWJ3Dnbhk-fLXQDkr4wEUgYKbbsTl (post deleted)

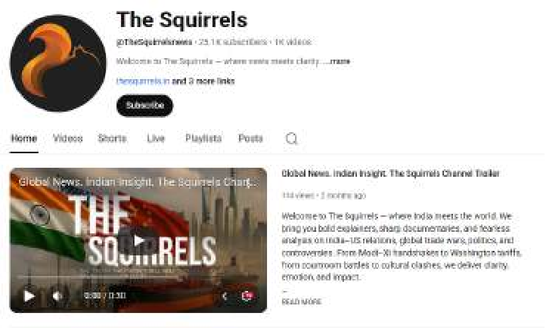
English translation:

Did he lift 12 million Argentinians out of poverty? And "send them into the most atrocious destitution"?

"The dream of owning a home is possible," Javier Milei stated, while "soup kitchens are overwhelmed, poverty is spiraling out of control, and more and more families are sleeping on the freezing sidewalks of Argentinian cities. The presidential phrase, celebrated as a permanent campaign slogan, does not reflect the reality of a country devastated by its own economic policies."

Milei claims that he "lifted 12 million Argentinians out of poverty," when what he actually did was "plunge the entire nation into the most brutal misery, surrendering economic sovereignty to the IMF and its international usurers, repeating a neighborhood phenomenon of global subservience where obeying foreigners is valued more than defending popular dignity."

"While the government repeats slogans and celebrates indicators that no worker sees reflected in their daily earnings, the population sees wages decimated, daily inflation soaring, and access to housing transformed into a propaganda fiction. Mortgage loans are now a luxury for millionaires, not a tool for Argentine families."



The Squirrels

@TheSquirrelsnews · 23.1K subscribers · 1K videos

Welcome to The Squirrels — where news meets clarity. ...more

thesquirrels.in and 3 more links

Global News. Indian Insight. The Squirrels Channel Trailer

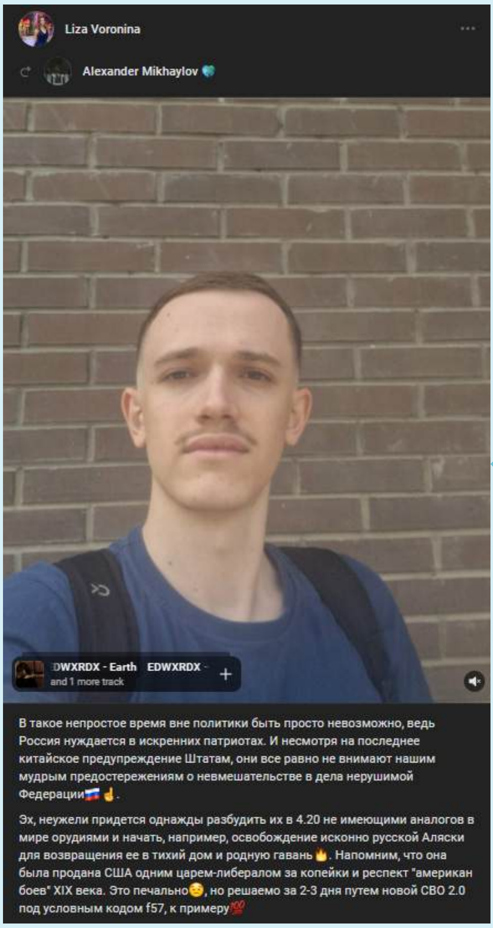
114 views · 2 months ago

Welcome to The Squirrels — where India meets the world. We bring you bold explainers, sharp documentaries, and fearless analysis on India–US relations, global trade wars, politics, and controversies. From Modi–Xi handshakes to Washington tariffs, from courtroom battles to cultural clashes, we deliver clarity, emotion, and impact.

READ MORE

On YouTube, several Indian political and news channels are followed by bots from our sample, including *ABP Live* and *The Squirrels*, which focus on politics, government affairs, and regional conflicts.

VKontakte continues to serve as an ecosystem for promoting pro-Kremlin narratives, where commercial bots amplify this agenda. VKontakte continues to serve as an ecosystem for promoting pro-Kremlin narratives, where commercial bots amplify this agenda. One illustrative, non-exhaustive example of this broader pattern involves a case where bots reposted a post that hints at the possible use of “unprecedented” weapons and fantasises about “liberating” Alaska in 2-3 days under a so-called “SVO 2.0,” while invoking historical revisionism about the sale of Alaska.



Liza Voronina

Alexander Mikhaylov

В такое непростое время вне политики быть просто невозможно, ведь Россия нуждается в искренних патриотах. И несмотря на последнее китайское предупреждение Штатам, они все равно не внимают нашим мудрым предостережениям о невмешательстве в дела нерушимой Федерации🇷🇺☝️.

Эх, неужели придется однажды разбудить их в 4.20 не имеющими аналогов в мире орудиями и начать, например, освобождение исконно русской Аляски для возвращения ее в тихий дом и родную гавань🔥. Напомним, что она была продана США одним царем-либералом за копейки и респект "американ боев" XIX века. Это печально😔, но решаемо за 2-3 дня путем новой СВО 2.0 под условным кодом f57, к примеру💯

English translation:

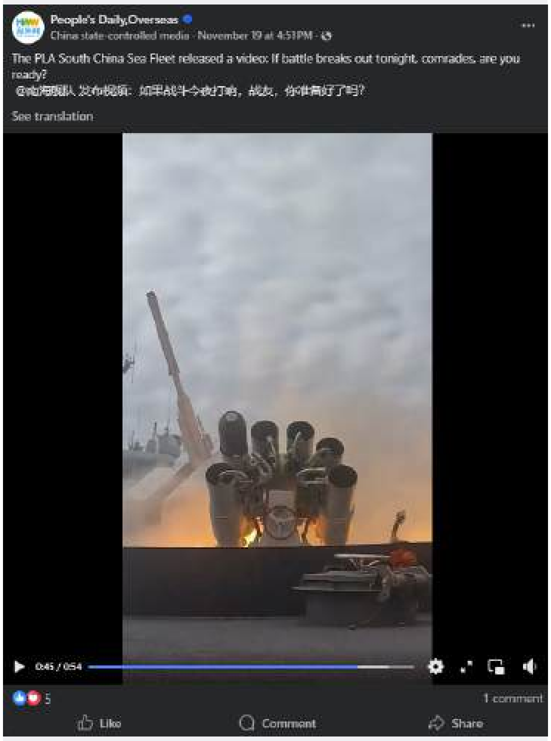
In such difficult times, it’s simply impossible to stay out of politics, because Russia needs sincere patriots. And despite China’s latest warning to the United States, they still aren’t heeding our wise warnings about non-interference in the affairs of the inviolable Federation. 🇷🇺☝️.

Oh, will we really have to wake them up one day at 4:20 with weapons unparalleled in the world and begin, for example, the liberation of the primordially Russian Alaska, returning it to its quiet home and native harbor. 🔥 *Let’s remember that it was sold to the United States by a liberal tsar for pennies and respect for the “American Boys” of the 19th century. This is sad,* 😟*, but it can be resolved in 2-3 days with a new SVO 2.0, codenamed f57, for example.* 💯

https://vk.com/id356957878?w=wall356957878_3286

Shifting focus towards amplifying China’s strength

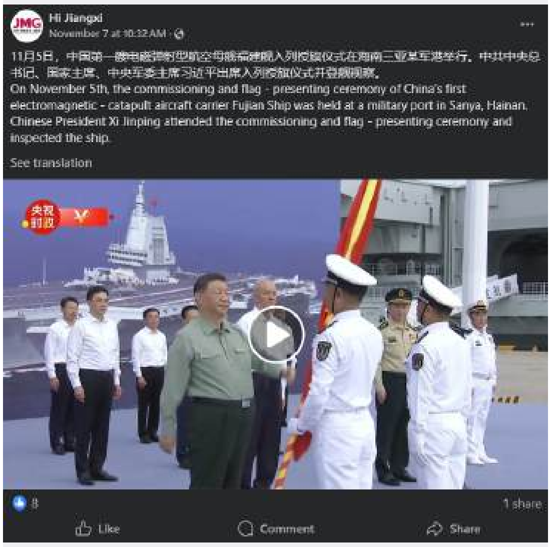
In addition, this year **we observed a broader shift in bot-promoted narratives from primarily local political content toward military-related themes**. Across Facebook, X, YouTube, and TikTok, we detected bot activity amplifying China-related military narratives, including content praising China’s military strength. While cross-platform volumes are not directly comparable due to platform-specific differences in data access and visibility, the presence of this theme across multiple platforms represents a notable departure from prior years in our experiments. On Facebook, for example, hundreds of identified bot accounts followed pages presenting themselves as Chinese news sources, including one labeled as “China state-controlled media.” These pages posted militarised content such as videos of missile launches, naval operations, the commissioning of new military vessels, and official state ceremonies attended by senior Chinese officials.


People's Daily,Overseas
China state-controlled media · November 19 at 4:51PM ·

The PLA South China Sea Fleet released a video: If battle breaks out tonight, comrades, are you ready?
@南海舰队 发布视频：如果战斗今夜打响，战友，你准备好了吗？

See translation

https://www.facebook.com/Haiwainet

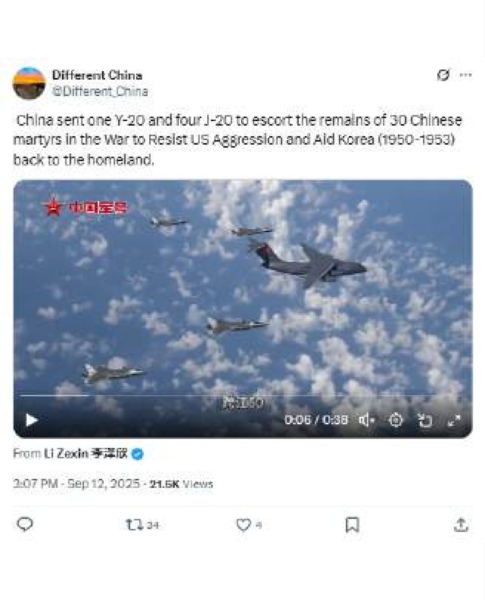

Hi Jiangxi
November 7 at 10:32 AM ·

11月5日，中国第一艘电磁弹射型航空母舰福建舰入列授旗仪式在海南三亚某军港举行。中共中央总书记、国家主席、中央军委主席习近平出席入列授旗仪式并登舰视察。
On November 5th, the commissioning and flag - presenting ceremony of China's first electromagnetic - catapult aircraft carrier Fujian Ship was held at a military port in Sanya, Hainan. Chinese President Xi Jinping attended the commissioning and flag - presenting ceremony and inspected the ship.

See translation

https://www.facebook.com/hijiangxi

On X, alongside other pro-China content, bots amplified posts portraying the Chinese military as superior to the United States.



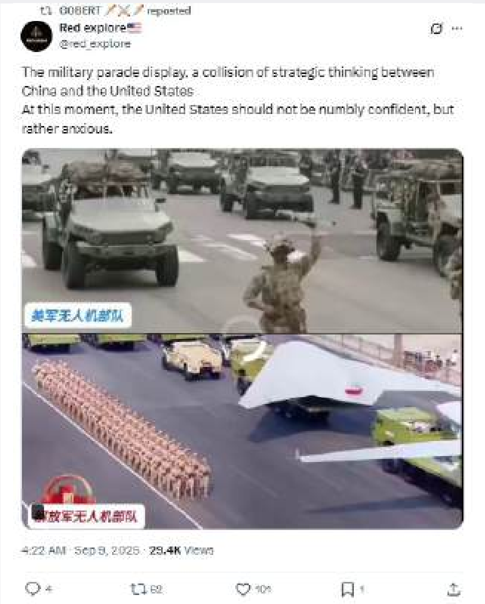
Different China
@Different_China

China sent one Y-20 and four J-20 to escort the remains of 30 Chinese martyrs in the War to Resist US Aggression and Aid Korea (1950-1953) back to the homeland.

From Li Zexin 李泽欣

3:07 PM · Sep 12, 2025 · 21.5K Views

https://x.com/Different_China/status/1966473841154101709



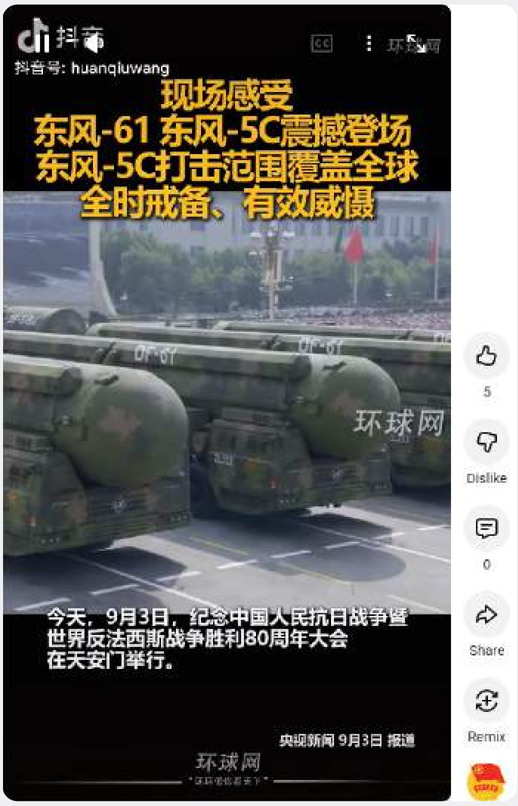
GOBERT reposted

Red explore
@red_explore

The military parade display, a collision of strategic thinking between China and the United States
At this moment, the United States should not be numbly confident, but rather anxious.

4:22 AM · Sep 9, 2025 · 29.4K Views

https://x.com/GOBERT1384477/status/1966335852440859120



抖音号: huanqiuwang

现场感受
东风-61 东风-5C震撼登场
东风-5C打击范围覆盖全球
全时戒备、有效威慑

今天，9月3日，纪念中国人民抗日战争暨世界反法西斯战争胜利80周年大会在天安门举行。

央视新闻 9月3日 报道

On YouTube, bots from our sample followed a channel that shared Chinese patriotic content. The channel featured military parades and missile demonstrations, including the Dongfeng series of long-range ballistic missiles, highlighting China's military capabilities, as well as patriotic videos and messages promoting national unity under the "One China" policy.

English translation:

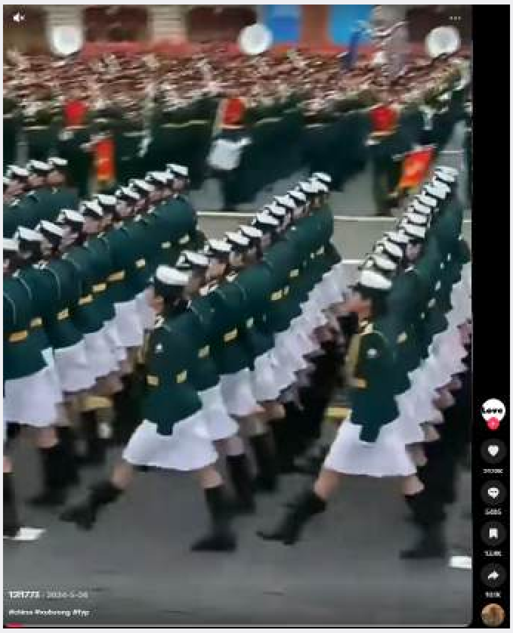
Impressive scene!

Dongfeng-61 and Dongfeng-5C make a stunning appearance.

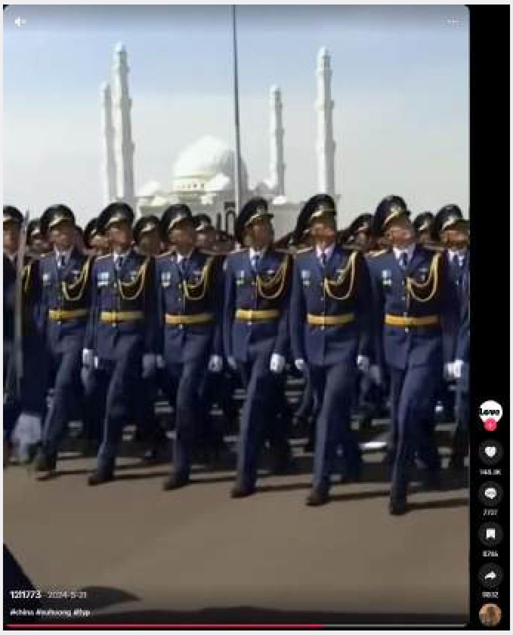
The Dongfeng-5C has a strike range covering the entire world.

Always on standby, providing effective deterrence.

On TikTok, bots reposted content from the account glorifying both the Chinese and Russian armies at the same time. The content focused on visual aesthetics with synchronised marching, music, and discipline, portraying both armies as powerful and united.



https://www.tiktok.com/@12l1773/video/7372501811153898785



https://www.tiktok.com/@12l1773/video/7371360375935225120

Our monitoring revealed that Instagram was the only platform where no bots were detected amplifying political content; this activity was observed on all other platforms included in the experiment, engaging with political and military themes. While bot amplification of such content is not new, a significant trend emerged this year, following the super-election year: a notable increase in narratives focusing on Chinese military strength. This shift suggests a potential increase in demand for the amplification of these narratives; while the experiment cannot establish direct causality, this pattern is consistent with broader changes in the information environment observed during the reporting period. Accordingly, these findings should be interpreted within the broader context of geopolitical events, kinetic activities, and evolving narratives in the digital domain.

Addressing the scale of the market: Assessment by The Financial Intelligence Unit of Latvia

Traceability observation

Complete traceability of cryptocurrency transactions remains elusive, despite the use of advanced analytical tools and blockchain tracing technologies. Our analysis of 10 transactions showed that only four, where funds were sent directly to the social media marketing panel service provider's wallet, could be reliably identified. While based on a limited sample, these observations provide indicative insight into service-provider transaction patterns within the constraints of custodial payment infrastructures. This proportion is nonetheless sufficient for a preliminary assessment of the service provider's operational scope, transaction dynamics, and financial indicators, enabling the formulation of substantiated conclusions for this study.

In contrast, the remaining 6 transactions could not be fully traced end-to-end. The funds for these transactions were routed through virtual currency payment platforms, including the hot wallets of virtual asset service providers (VASPs). This routing leads to the commingling of funds from multiple users at these addresses, making it impossible to trace individual transaction flows using only on-chain data. This limitation is typical of custodial systems, as exchanges routinely pool client deposits and redistribute them internally, which breaks the observable chain of on-chain attribution.

Description of transactions

We selected commercial manipulation providers across a broad geographic spread-including North America, Western and Northern Europe, Eastern Europe, Russia, and parts of Asia - and assigned the following reference labels: US1, US2, UK1, UK2, IT, EST, RU1, RU2, CHN, and IND. These labels serve as anonymised identifiers for analytical reference and do not necessarily correspond to the verified physical location or legal identity of the specific providers.

The social media experiment comprised 10 cryptocurrency transactions executed on a single day, with a cumulative notional value of approximately USD 230. Three assets were used: Bitcoin (BTC), Polygon (MATIC), and Tether (USDT). Transactions were routed through two virtual asset service providers, Binance and Bybit. All transfers originated from exchange-hosted custodial hot wallets, which are typically used for rapid client transactions and liquidity management.

Two transfers to CHN and IND used BTC and were initiated from Binance. Because they came from a Binance hot wallet, they may reflect Binance's practice of aggregating unrelated user withdrawals into one outbound transaction to improve processing efficiency and reduce fees. The BTC was sent to two addresses linked to Cryptomus. After receipt, Cryptomus co-spent these funds alongside other inputs and forwarded them to outputs without clear attribution.

Four transfers to US2, IT, EST, and RU2 used MATIC and were also initiated from Binance. Funds were sent to four single-use addresses and then forwarded onward. In two cases, they went to an address attributed to Cryptomus, and in two cases to an address attributed to Heleket, a higher-risk exchange due to limited regulatory oversight and unclear KYC and AML controls. Further tracing is not possible using on-chain data alone once funds enter Cryptomus or Heleket wallets, since internal movements are not visible on the public blockchain and would require provider cooperation or legal assistance.

Four transfers to US1, RU1, UK1, and UK2 used USDT and were initiated from a Bybit hot wallet, then sent to Binance deposit addresses. Analytical tools linked two recipient addresses to service providers UK1 and UK2, enabling a partial estimate of their transaction volume based on observable activity. The other two recipient addresses, tied to US1 and RU1, were not attributed by tools, but their use in the experiment and the observed flow patterns suggest they are dedicated intake addresses for those providers, allowing a reasonable partial estimate of their handled volume.

Time vs. volume

The transaction volume of a given address can be assessed over an extended period beginning from the address's first on-chain activity up to the end of the observation window used in this study (November 2025).

As shown in the visualisation (Figure 13, page 29), blockchain analytical tools enable a detailed examination of both incoming and outgoing transactions associated with the address. However, because cryptocurrency address generation is rapid and practically effortless, identifying the full scope of a service provider's on-chain activity remains challenging. This limitation reflects incomplete address attribution rather than a lack of analytical tooling, as not all addresses associated with a given provider can be reliably identified.

The considerable volume of USD-denominated transactions processed by the chosen providers suggests a significant volume of services were rendered. **Notably, Russian provider RU1 demonstrates a very high volume of transaction activity in 2024, which significantly decreased following the end of the *election year.***

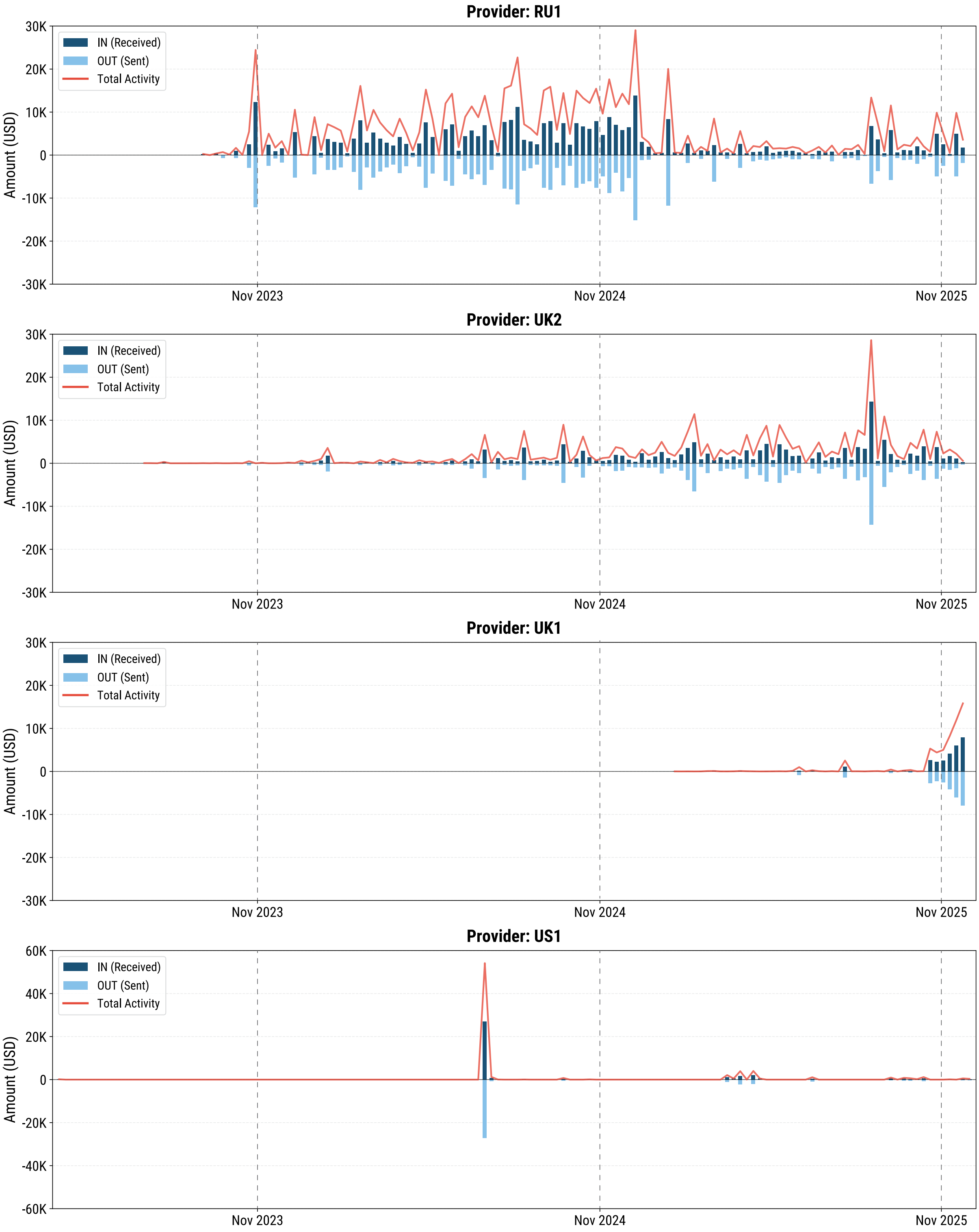


FIGURE 13. Transactional Flow Over Time. This figure presents a timeline illustrating the incoming and outgoing transactions processed by four distinct manipulation providers. Values on the Y-axis are thresholded to 30,000 except for US1 provider. The red line represents IN + |OUT| transactions and is interpreted as total activity.

Estimated market volume

The following section presents a general analysis, based on the available data, of the total transaction volume (in USD) received by each identified service provider since the initial use of the respective wallet addresses. Overall assessment is provided below:

US1 Analysis indicates that the service provider's Binance deposit address was first used on April 5, 2023, with the most recent transaction recorded on October 18, 2025. As the latest transaction occurred less than a month ago, it can reasonably be concluded that the address remains active.

A total of 42 incoming transactions were identified for this address, amounting to an approximate value of **USD 36,152**. Over the period from April 2023 to October 2025 (approximately 30 months), this corresponds to an average monthly transaction volume of about USD 1,205.

RU1 Analysis indicates that this service provider's Binance deposit address was first used on 5 September 2023, with the most recent transaction recorded on 30 October 2025[1]. As the latest transaction occurred on the date of analysis, it can be concluded with a high degree of confidence that the address remains active.

A total of 681 incoming transactions were identified, amounting to approximately USD 265,261. Over the period from September 2023 to October 2025 (approximately 26 months), this corresponds to an average monthly transaction volume of about USD 10,202. Incoming transactions primarily originated from Cryptomus, which accounted for the largest share of volume, as well as Heleket.

The involvement of Heleket is notable from a regulatory perspective. In addition, expert input suggests that the RU1 operator is based in Russia and is likely to use a Binance wallet for receiving funds.

Council Regulation (EU) No. 833/2014, Article 5b(2), prohibited up to 23 October 2025 the provision of crypto-asset wallet, account, or custody services to Russian nationals, persons residing in Russia, or entities established in Russia. As of 24 October 2025, the provision was broadened and now prohibits providing, directly or indirectly, (a) crypto-asset services as defined in Regulation (EU) 2023/1114 (MiCA), (b) issuance or acquisition of payment transactions or payment initiation services, and (c) issuance of electronic money, to the same categories of Russian persons or entities.

Accordingly, if the Binance legal entity servicing this wallet were subject to EU restrictive measures, and the RU1 operator/wallet holder qualifies as a Russian national/resident or a Russia-established entity, Binance's provision of these services would fall within the scope of Article 5b(2) and could constitute a sanctions breach. Determining this conclusively requires confirmation of (i) the specific Binance entity providing the service and (ii) the legal status/nationality or place of establishment of the wallet holder. Binance maintains several EU-linked registrations and authorisations, meaning EU restrictions can be relevant to parts of its operations.

Given the frequency and scale of transactions, the RU1 operator appears to operate on a comparatively large scale, maintaining substantial and consistent transaction flows. It is also reasonable to assess that the entity relies on multiple cryptocurrencies, wallet addresses, and service providers to facilitate payment processing.

UK1 Analysis indicates that this service provider’s Binance deposit address was first used on January 26, 2025, with the most recent transaction occurring on October 29, 2025. As the latest transaction occurred a day before the date of analysis, it can be concluded with a high degree of confidence that the address remains active.

A total of 77 incoming transactions were recorded, amounting to an approximate value of **USD 7,808**. Over the period from January to October 2025 (approximately 10 months), this corresponds to an average monthly transaction volume of approximately USD 781.

UK2 Analysis indicates that this service provider’s Binance deposit address was first used on July 7, 2023, with the most recent transaction recorded on October 30, 2025. As the latest transaction occurred on the day of analysis, it can be concluded with a high degree of confidence that the address remains active.

A total of 1,307 incoming transactions were recorded, amounting to an approximate value of **USD 123,714**. Over the period from July 2023 to October 2025 (approximately 28 months), this corresponds to an average monthly transaction volume of approximately USD 4,418.

The substantial and persistent market for social media manipulation persists due to gaps in monitoring, disruption, regulation, and accountability. While transparency reports confirm that social media platforms do actively monitor these services and detect inauthentic accounts, these efforts would need to be elevated to a strategic level to effectively disrupt foreign interference worldwide.

AI perspective

Testing AI-enabled orchestration

> Automated content publishing across all targeted platforms is completely feasible.

We developed a workflow to automatically generate and publish content using commercial content orchestrators. Following initial configuration, content generation and publication proceeded without further human intervention. We designed prompts describing the type of post to be generated and stored them in a Google Sheets spreadsheet. ChatGPT (GPT-4o, via API) produced the text for each post. For images and videos, we used the Freepik’s API, which provided a pre-trained character model that remained consistent across all platforms.

The chosen orchestrator offered built-in integrations with social media platforms, enabling the system to connect with inauthentic accounts and automatically publish the generated posts. For videos, the workflow used an additional prompt to convert the image into a short video using AI tools before uploading. The entire process functioned as a fully closed-loop, covering text generation, visual creation, post publication, and results logging in the spreadsheet without any manual intervention.

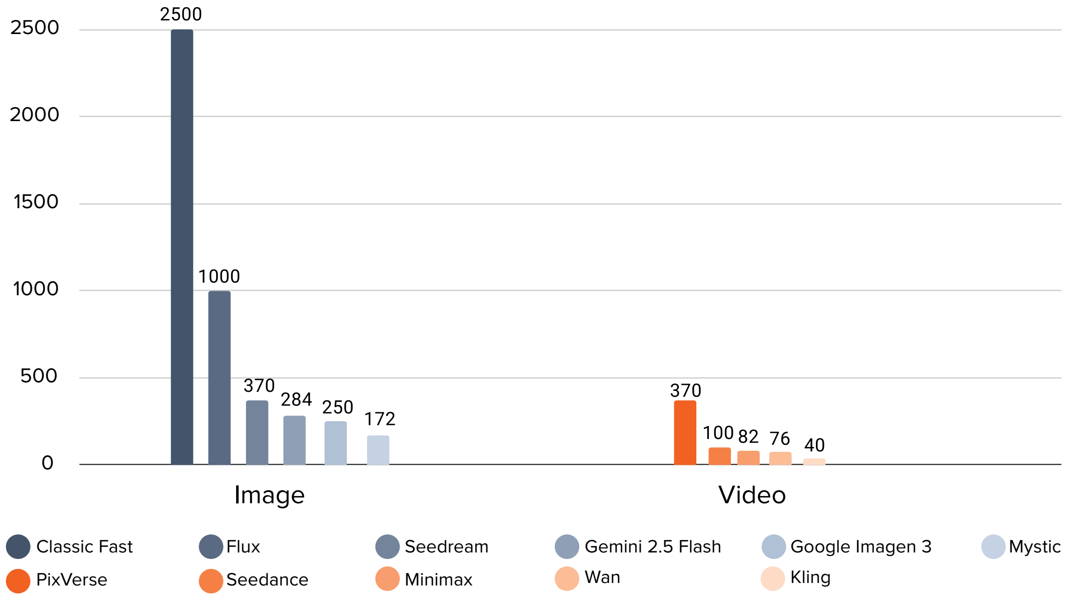


FIGURE 14. What volume of AI-generated content is attainable for a budget of 10 EUR?

Our findings (Figure 14) suggest that the large-scale dissemination of AI-generated content is relatively cheap and easily accessible. For example, for €10, it is possible to generate hundreds of videos and thousands of images. This illustrates the low barriers and limited effort required to conduct manipulative campaigns using AI-generated content.

The established and tested content distribution workflow, leveraging existing solutions, has proven successful. We have validated this automated system across all experimental platforms, where content was published without any manual intervention. The outcomes suggest that extending this automation to additional platforms is likely technically feasible, though such scaling remains subject to the specific architectural constraints of each environment.

New frontier? The AI-enabled sophistication of bots – Cyabra analysis

Overview

In the series of Social Media Manipulation Experiments, we focus on commercial manipulation that enables amplification of posts using spam bots. In today's context, spam bot identification is only a part of the problem due to AI-enabled bot behaviour sophistication that enables more targeted and covert operations. In this regard, later in this report, we provide an assessment of the evolution of inauthentic user (bot) sophistication.

First, we established a "legacy bot" baseline using Cyabra's 2018–2023 datasets acquired through their proprietary system in combination with identified inauthentic accounts in our archives, characterised by repetitive behaviour, uniform language, centralised control, and predictable engagement. For this study, we also collected and normalised updated 2024-2025 datasets from X, Facebook, TikTok, and VK, enabling direct comparison across time and platforms. **Cyabra's analytical models were applied to contrast historical and current features-examining linguistic variability, synchronisation patterns, use of AI-generated multimodal content, and engagement anomalies.** This revealed clear differences between conventional spam bots and more adaptive, human-like automated accounts. These findings were validated through manual review by Cyabra's analyst team to ensure the behavioural differences are genuine. Overall, the **comparison shows that modern bots now employ context-aware content, sustained low-volume activity, and more organic, decentralised coordination, indicating a significant rise in operational sophistication.**

Findings

Simpler operations depend on standalone posting: fake profiles publish material on their own pages, boosting it with hashtags, and interacting mainly within closed, inauthentic networks. While this approach successfully increases volume, it often fails to connect with genuine audiences.

In recent and more sophisticated campaigns, hostile disinformation operators have shifted toward embedding their messages directly inside authentic conversations. Instead of relying on self-contained posting loops, fake profiles now target high-visibility content created by influencers, journalists, and public figures, placing crafted comments beneath posts that already attract genuine engagement. This approach allows campaigns to blend into organic discourse and appear far more credible. This shift in sophistication is due to AI-enabled automation capabilities. **Automated systems identify relevant influencer posts, track trending discussions, and generate context-aware comments in multiple languages, significantly reducing the detectable traces of coordination.** Generative models produce short, natural-sounding messages that match the tone and topic of the original post, giving fake profiles an authenticity they previously lacked.

In this study, we seek to identify the main indicators one must take into account when building a Coordinated Inauthentic Behaviour (CIB) detection pipeline. From the findings, we have identified the following changes in behaviour: **reduced interaction among fake**

accounts and increased engagement with verified or high-follower profiles; strong contextual alignment between fake comments and the host post; declining reliance on hashtags; and greater message exposure achieved through the influencer's audience rather than artificial amplification. Overall, the strategy has moved from visibility through sheer volume to influence through precise placement. By entering credible comment spaces, even a small number of AI-generated posts can shape perceptions more effectively than large, isolated posting networks for example, fake accounts commenting under journalists' election-related posts to subtly promote or undermine political figures.

As an example Cyabra's analysis of online discourse surrounding the September 2025 Russian drone incursion into Polish airspace, illustrates how AI has transformed modern disinformation efforts, enabling rapid creation, coordination, and amplification of false narratives. Of the 3,622 analysed profiles, 22% were identified as inauthentic-far above typical baselines-actively promoting narratives that deflected blame from Russia, undermined trust in Polish leadership, and downplayed the severity of the incident. These fake accounts represented a new generation of highly sophisticated profiles that used AI-generated imagery, credible local naming conventions, and professional-looking media personas to appear authentic. Their behaviour showed coordinated timing, stylistic variation, and multilingual, context-aware engagement, all supported by AI-driven text generation. **This case study is indicative of an emerging behavioural pattern observed in this analysis, rather than a comprehensive assessment of all contemporary disinformation activity. Beyond classical mechanical automation (spam bot behaviour), clever implementation of AI now facilitates psychologically realistic manipulation by enabling fake profiles to seamlessly integrate into authentic online discourse, mimic human behaviour, and exploit genuine emotional reactions.**

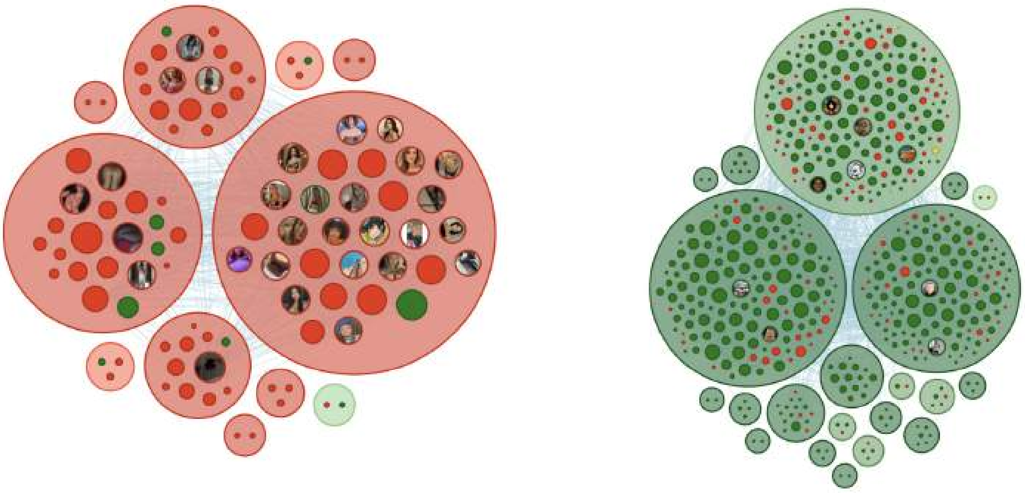


FIGURE 15. The visual presents a comparison between clusters of inauthentic accounts (bots, colored red) and real users (colored green).

The cluster (Figure 15) on the left represents the behaviour of the amplifier bots. These accounts form dense connections with one another, creating a large, closed community composed almost entirely of inauthentic profiles. On the right, the cluster illustrates a newer generation of fake accounts with more advanced and aligned with the behavioural traits described earlier. Here, the bots establish connections not only with one another but also with authentic users. They integrate into real conversations and existing communities, blending in almost organically.

Conclusions

Overall improvements: Social media platforms have demonstrated progress in combating spam-related activities. Although the rate of blocking new account creation remains consistent with the prior assessment period, there has been an improvement in the removal of existing accounts and the truncating of spam-related activity.

Scaling problem: Manipulation is still easy to make and cheap even though we scaled up an experiment. While the cost of manipulation remains low, the effectiveness of removal and moderation efforts is increasing. Consequently, adversaries must now expend greater resources to achieve the same impact they previously accomplished more cheaply.

Shifting focus of spam bot influence: Following the 2024 election year, the focus of bot-driven content has emerged from political matters to military topics. Specifically, our analysis of sampled bots indicates significant amplification of pro-Chinese content.

Social media advertisement manipulation is affordable despite being significantly more expensive than regular manipulation. The manipulation market exists and buying accounts able to run ads remains easy. Once purchased, simple AI-enabled orchestration services allow distributing content in cross-platform format.

Cryptocurrency as a gateway to manipulations: Commercial manipulation providers use cryptocurrency as their main payment mechanism because it is fast, cross-border, and hard to police. They typically route customer funds through custodial wallets and high-risk exchanges, which gives them a resilient and low-visibility financial backbone. A key tactic is sending deposits into VASP hot wallets like Cryptomus and Heleket, where many users' funds are mixed together. This commingling breaks the on-chain attribution trail, so full traceability is rarely possible; in the experiment, only four out of ten transactions could be reliably tracked end-to-end. Even so, the deposit addresses that could be linked to providers show steady, high-volume activity over time, confirming that the manipulation market is large and ongoing. The money involved is meaningful: RU1 received about USD 265,261 and UK2 about USD 123,714 between September 2023 and October 2025, which implies frequent service delivery and sustained demand. There are also legal and regulatory risks, especially when suspected Russia-based operators cash out through major exchange custody such as Binance, potentially triggering EU sanctions compliance concerns under Council Regulation (EU) No. 833/2014, Article 5b(2). Taken together, cryptocurrencies' current payment infrastructure enables these providers to operate at scale with limited oversight, strengthening the case for deeper cooperation with financial intelligence units and VASPs to identify actors and restrict their capacity.

Levels of sophistication: AI-driven networks are now employing sophisticated fake profiles that achieve psychological realism and behavioural convergence. These synthetic actors appear genuinely human by utilising credible visuals, localised language, and adaptive timing. Crucially, these networks have changed their strategy from simply pushing content to embedding themselves directly within communities. By earning trust and performing authenticity, they are able to blend into everyday dialogue, gaining access to where opinions are formed. This allows them to subtly steer sentiment, either by deepening divides or by promoting targeted viewpoints, from a position of earned trust within the community, rather than through sheer volume from an external source. The increasing sophistication of bot activity now poses a significant challenge to platforms attempting to combat inauthentic manipulation. Consequently, this introduces new layers of complexity to our vulnerability assessments. The rise of sophisticated digital operations available as a service presents a significant challenge to our online environment.

Implications for the digital space:

Threat to authenticity: The commercial availability of these sophisticated operations means they are accessible globally, posing a massive threat to the integrity and genuineness of online conversations.

A new challenge for information integrity: Social media platforms continue to struggle with basic spam enforcement, while increasingly sophisticated bot operations are pushing content moderation into a far more complex and contested domain, with direct implications for the protection of information integrity and free expression.

Recommendations

Prioritise behavioural detection: Shift focus from mere textual analysis to detecting cross-platform behavioural synchronisation. Key indicators are coordinated patterns in timing, tone, and relational dynamics, as these, rather than simple copy-paste text, now signal sophisticated AI influence.

Adopt continuous, contextual monitoring: Move beyond traditional “campaign thinking” to a model of continuous monitoring. This involves tracking how narratives evolve over time and pinpointing where they penetrate genuine dialogue spaces.

Map conversation-level influence: Analyse conversations, not just individual posts, to identify fake profiles embedded within authentic threads and influencer communities. This conversation-level mapping enables earlier detection of “in-conversation” manipulation.

Follow the money: Financial transactions should be a key component of influence campaign analysis. Significant pro-active financial investment into social media manipulation services can serve as an indicator of broader, strategic online influence operations.

Play the red-team and develop disruption mechanisms: To proactively address vulnerabilities, it is essential to first understand the gray market and accurately identify and assess manipulation services. Subsequently, policymakers and social media platforms must be promptly informed of these vulnerabilities. Furthermore, companies offering manipulation services should be sanctioned. Proactive disruption strategies must be continuously updated with the most current research and findings.

Endnotes

[1] An engagement delivery refers to any instance of interaction or activity specifically performed on a social media platform.

[2] An inauthentic comment is a type of engagement produced by a fake, automated, or coordinated account intended to deceive others or manipulate platform algorithms rather than express a genuine, independent opinion.

[3] The "expected number" represents the total volume of comments explicitly purchased from the manipulation service provider (the 100% baseline), meaning the 340% figure reflects an actual delivery of more than triple the contracted quantity within 72 hours.

[4] Soper, S., & Dave, P. (2025, December 15). Meta tolerates rampant ad fraud from China to safeguard billions in revenue. Reuters. ***www.reuters.com/investigations/meta-tolerates-rampant-ad-fraud-china-safeguard-billions-revenue-2025-12-15***

[5] Dek, A., Kyrychenko, Y., van der Linden, S., and Roozenbeek, J., "Mapping the online manipulation economy: A market perspective on digital manipulation may help improve online trust and safety", Science, 2025. ***https://doi.org/10.1126/science.adw8154***

[6] COTSI; cotsi.org

[7] This assumption reflects the timeline of the experiment.

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