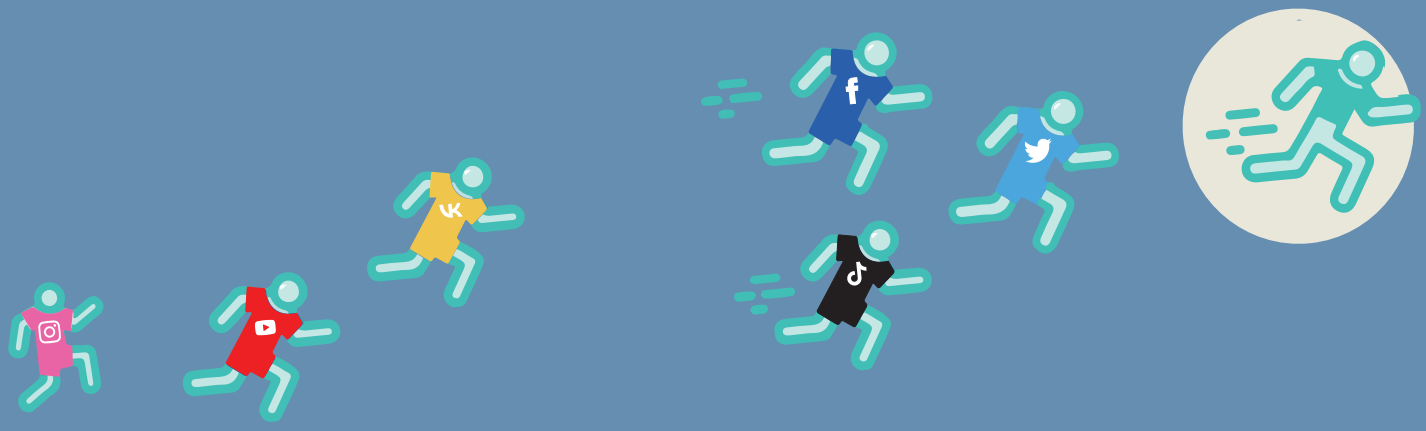
ISBN:



SOCIAL MEDIA MANIPULATION 2021/2022:
ASSESSING THE ABILITY OF SOCIAL MEDIA COMPANIES TO COMBAT PLATFORM MANIPULATION

PREPARED BY THE
NATO STRATEGIC COMMUNICATIONS
CENTRE OF EXCELLENCE

ISBN:
Project manager: Rolf Fredheim
Authors: Sebastian Bay, Rolf Fredheim
Research and data analysis: Trementum Analytics
Copy Editor: Monika Hanley
Design: Linda Curika

This report was completed in April 2022, based on an experiment that was conducted from September to November 2021. Discussions with social media companies regarding preliminary results took place in January and February 2022.

Trementum Analytics is an analytics, software development, and data visualisation company based in Kharkiv, Kyiv, and Cambridge.
Website: www.trementum.net
Email: contact@trementum.net

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

EXECUTIVE SUMMARY

Introduction

The ongoing war in Ukraine has shown the importance of being able to defend and influence the information environment in order to win a modern conflict. Coordinated social media manipulation campaigns continue to be an important tool for adversaries.

Therefore, assessing the abilities of social media companies to protect their platforms against manipulation continues to be important for understanding how well our societies are able to protect our information environment against antagonistic threats.

To further our understanding of platform manipulation, we re-ran our groundbreaking experiment to assess the ability of social media companies to counter platform manipulation. This year, we added a sixth social media platform, VKontakte, to our experiment.

The Experiment

To test the ability of social media companies to identify and remove manipulation, we bought inauthentic engagement on 46 posts on Facebook, Instagram, Twitter, YouTube, TikTok, and VKontakte ("the platforms"), using three high-quality Russian social media manipulation service providers.

For €279, we received inauthentic engagement in the form of 1,385 comments, 13,859 likes, 93,009 views, and 5,808 shares on Facebook, Instagram, YouTube, Twitter, Tiktok, and VKontakte, enabling us to identify 9,846 accounts used for social media manipulation. Of the 114,061 fake engagements purchased, more than 96 per cent remained online and active after four weeks, and, even after discounting fake views, more than 86 per cent of the 21,052 other fake engagements remained active after a month. It is still easy to buy fake engagements on social media platforms.

While measuring the ability of social media platforms to counter platform manipulation, it became clear that some of the platforms had made important improvements, but that the overall picture shows a continued inability to combat manipulation.

How easy is it to create a fake account on social media platforms? We conclude that it is still possible to buy many fake accounts at low cost from the blossoming fake account industry. We see no indications that this is changing. To date, none of the platforms' actions have materially changed the functioning of the manipulation industry.

How long does it take for half of all identified inauthentic accounts to be removed? Only Twitter has managed to consistently, on a year-to-year basis, reduce the half-life of the account population engaged in inauthentic behaviour. VKontakte quickly identifies and removes a majority of accounts engaged in inauthentic behaviour. On Facebook, Instagram, TikTok, and YouTube, very few, if any, accounts conducting inauthentic engagement are identified and removed from the platforms.

How quickly do the platforms remove inauthentic engagement: TikTok performed significantly better than in 2020 and went from removing the least to removing the most activity, probably as a result of the platform's strengthened counter-abuse efforts implemented during the past year. YouTube performed slightly better, while Twitter, Instagram, and Facebook performed worse compared to last time. VKontakte performed the worst, yet only marginally worse than Facebook and Instagram.

The price of a basket of manipulation: we compared 100 likes, 100 comments, 100 followers, and 1000 views from six Russian manipulation service providers to arrive at a median price for 2021 and compared it to the previous years' assessments, as well as to historical data. The results indicate that there is no significant shift in the cost of social media manipulation. It remains cheaper to buy automatic manipulation, such as views and likes, while more labour-intensive manipulation, such as comments, are several times more expensive.

It is interesting to note the significant price difference for followers between YouTube, Facebook, and Twitter, on the one hand, and Instagram, TikTok, and VKontakte on the other, with the latter group being between five and fifteen times as expensive. For comments, Instagram and VKontakte stand out as especially cheap. For the third year in a row, Twitter is the most expensive platform to manipulate and Instagram is the cheapest.

In 2021, on Instagram, €10 was enough to purchase almost 100,000 fake views, 25,000 fake likes, 1,000 fake comments, or 15,000 fake followers.

Speed of manipulation: on average, 20 per cent of all manipulation was delivered within an hour. After six hours, more than 30 per cent of the manipulation had been delivered on all the social media platforms. This indicates it remains possible to manipulate messaging about current events using commercial service manipulation providers, and—unfortunately—manipulation is getting faster rather than slower.

Removal of inauthentic accounts reported to the platforms: we observed that reporting an account still does not lead to that account being blocked or suspended, even if the reported account is known to have engaged in inauthentic activity. As in previous years, we conclude that reporting and moderation mechanisms must be improved so that a larger share of accounts flagged as inauthentic are acted upon, even if they are reported by a single user. It is

problematic that inauthentic accounts, even when reported as such, typically escape sanction.

Comparing the transparency of the platforms: we noted that, in a stunning disclosure, TikTok reportedly prevented 16.6 billion fake likes in a single quarter. Currently, no other platform reports the number of fake likes; one can only imagine what the corresponding figures might be. In Q3, the platforms in total reported that more than 22 billion fake engagements or fake accounts were prevented or removed, reflecting the enormous scale of the problem.

Conclusions

The most important insight from our study continues to hold: there is a significant difference among platforms in their ability and willingness to counter manipulation of their services. The effort social media companies put into countering abuse pays off and creates a more secure platform.

TikTok showed significant improvement across all the areas measured in our experiment and stands out as the most improved platform. Twitter remains the industry leader in 2021, with TikTok and Facebook close behind. Despite notable improvements by some, none of the six platforms we studied are doing enough to prevent manipulation of their services. The manipulation service providers are still winning the digital arms race.
It is still easy, cheap, and quick to manipulate social media platforms.

Based on our experiment, we continue to recommend that governments introduce measures to:

- Increase transparency and develop new safety standards for social media platforms.
- Establish independent and well-resourced oversight of social media platforms.
- Increase efforts to deter social media manipulation.
- Continue to pressure social media platforms to do more to counter the abuse of their services.

> There is a significant difference among platforms in their ability and willingness to counter manipulation of their services. The effort social media companies put into countering abuse pays off and creates a more secure platform.

INTRODUCTION

In the wake of the Russian invasion of Ukraine, Facebook, Twitter, and Google have all reported that they have identified and removed Russian and Belarussian disinformation networks on their platforms.[1] Facebook reported that one such attempt to undermine confidence in the Ukrainian government used fake accounts and inauthentic personas across a range of social media services including Facebook, Instagram, Twitter, YouTube, Telegram, Odnoklassniki, and VKontakte (VK). The inauthentic personas used AI-generated pictures to appear more believable as they attempted to develop the personas that ultimately pushed pro-Kremlin narratives about Ukraine.[2]

The Kremlin has long sought to foment unrest and dissent in Ukraine to gain supporters for its agenda globally and to suppress dissent among domestic audiences. Before the invasion, there were intense efforts by the Kremlin to fabricate[3] a pretext for the invasion and undermine the support for the Ukrainian government by promoting narratives about ongoing atrocities against the population in Donbas.[4] None of these elaborate efforts involving explosions, corpses, drone and bodycam videos, and disinformation established a casus belli for the Kremlin. Elliot Higgins, the founder of Bellingcat, expressed surprise that the quality of the Kremlin's efforts had deteriorated, calling them 'dumb and lazy'.[5]

The Kremlin's covert inauthentic coordinated behaviour on social media in support of the Kremlin's objectives in Ukraine also seems to have achieved little impact so far and their overt activities through state media and other affiliated structures have faced unprecedented counteractions from regulators and social media companies alike.[6]

It is still far too early to tell if the Kremlin has indeed failed to plan the information dimension of the military operation. It is possible the efforts were ineffective against non-Russian audiences to begin with, or that they were effective, but the open source community has not yet been able to identify the successes. In the early stages of the war, the Kremlin's messaging has rotated between outlandish notions and conspiracy theories at a dizzying pace. This may be an attempt to try different messaging in search of something that works. In the past, we have seen a similar response to unpredicted crises—in particular, the downing of the passenger jet MH17 or the protests following the disputed Belarussian elections in 2020.

As the war continues, the Kremlin's narratives will settle down. The conflict will continue also online and it will likely take time before we are ultimately able to assess the effectiveness and long term effect of the Kremlin disinformation machinery.

The coming weeks and months will be an important litmus test for the social media platforms and their ability to identify and counter inauthentic coordinated activities on their platforms. We will be following their activities and results closely to deepen our assessment of their abilities to counter platform manipulation and abuse.

Two years ago, the NATO StratCom Centre of Excellence carried out a groundbreaking experiment to assess the ability of social media companies to counter the malicious use of their services. We have shown that an entire industry has developed around the manipulation of social media, and we have twice concluded that social media companies were experiencing significant challenges in countering platform manipulation. Recent events also indicate that the enforcement of the policies of social media companies aimed at countering platform manipulation—removing content which violate the platforms community standards, as well as labelling state-sponsored media—have been inconsistently enforced so far.[7]

It continues to be difficult to independently assess the effectiveness of the ability of social media companies to defend their platforms against antagonists seeking to exploit them. It remains important to evaluate how well social media companies are living up to their commitments, and to independently verify their ability to counter the misuse of their platforms.

Building on our previous work, we have rerun our experiment to assess the ability of social media companies to combat platform manipulation in 2021.

The Social Media Manipulation Industry

Many of the conclusions from our initial report, The Black Market for Social Media Manipulation,[8] and from the last two[9],[10] iterations of this report still hold true—the manipulation market remains functional and most orders are delivered in a timely and accurate manner. Social media manipulation remains widely available, cheap, and efficient, and continues to be used by antagonists seeking to influence elections, polarise public opinion, sidetrack legitimate political discussions, and manipulate commercial interests online.

The social media manipulation industry feeds the market for inauthentic comments, clicks, likes, and follows. Buyers range from individuals seeking to boost their popularity, to influencers gaming the online advertising system, to state-level actors with political motivations. Social media manipulation relies on inauthentic accounts that engage with other accounts online to influence public perception of trends and popularity. Some inauthentic accounts are simple, [ro]bot-controlled accounts without profile pictures or content, used only to view videos or retweet content as instructed by a computer programme. Others are elaborate 'aged' accounts with long histories meant to be indistinguishable from genuine users.

Bots are a very cost-efficient way of generating artificial reach and creating a wave of 'social proof', as typical users are more likely to trust and share content that has been liked by many others.[11] Bot-controlled accounts cost only a few cents each and are expected to be blocked relatively quickly. More elaborate inauthentic accounts require some direct human control and can cost several hundred dollars to purchase, often remaining online for years. Coordinated inauthentic engagement can also be achieved using authentic accounts, such as the campaign where Russian TikTok influencers were paid to spread pro-Kremlin narratives about the Russian invasion of Ukraine.[12]

Developments in 2022

A new ethnological study by Johan Lindquist provides an in-depth understanding of the social media manipulation industry from an insider's perspective. Over three years, Lindquist conducted 40 in-person and digital interviews with resellers of social media manipulation in eleven countries.[13] Lindquist describes a vast global industry. Information provided by Lindquist's informants indicate that there are more than 300 global providers of inauthentic accounts used by the social media manipulation industry. This industry consists of an untold number of resellers (likely many thousands) working at various levels in the industry—from large scale and highly automated, to manual one-man coffee-shop operations. Lindquist describes a highly connected, API-interlinked industry of sellers and resellers.

The six primary social media manipulation service providers we follow have all been operating for around ten years, and between them they claim to serve hundreds of thousands of customers around the world, executing thousands of orders a day with a staff typically ranging from 10 to 30 people per company. It is clear from our interaction with these large manipulation service providers that they continue to prosper and function as in previous years. None of the counteractions developed by social media platforms have fundamentally impacted their effectiveness or their prices in the last year. In fact, social media manipulation is, if anything, getting cheaper, faster, and more effective.

Three Insights

1. The scale of the industry is immense. The infrastructure for developing and maintaining social media manipulation software, generating fictitious accounts, and providing mobile proxies is vast. We have identified hundreds of providers. Several have many employees and generate significant revenues. It is clear that the problem of inauthentic activity is extensive and growing

2. During recent years, the manipulation industry has become increasingly global and interconnected. European service providers rely in particular on Russian manipulation software and infrastructure providers who, in turn, use contractors from Asia for much of the manual labour required. Social media manipulation is a global industry with global implications.

3. The openness of this industry is striking. Rather than lurking in a shadowy underworld, it is an easily accessible marketplace that most web users can reach with little effort through any search engine. They act in the open, yet thrive. Remarkably, social media companies have not found a way to put them out of business. In fact, manipulation service providers still advertise openly on major social media platforms and search engines.

Who We Are

The NATO Strategic Communications Centre of Excellence is a multinationally constituted and NATO-accredited international military organisation. We are not part of the NATO command structure and are not subordinate to any other NATO entity. Our strength is built on multinational and cross-sector collaboration of experts and analysts from the civilian, military, private, and academic sectors, and from the use of modern technologies and virtual tools for analysis, research, and decision making.

Since the centre was founded in 2014, we have studied social media manipulation as an important and integral part of the influence campaigns directed by malicious state and non-state actors against the Alliance and its partners. We have published numerous reports on this and related topics, including our quarterly Robotrolling[14] report, trend analyses, case studies, in-depth research, and armed forces assessments.

The malicious use of social media is a tool of choice for actors conducting influence activities against EU and NATO interests. Bolstering our collective resilience requires a deeper understanding of this problem so that we can perform accurate analyses, create early detection protocols, and establish effective prevention measures. This will be possible only if we identify and address the vulnerabilities of social media platforms.

The scale of the industry is immense. The infrastructure for developing and maintaining social media manipulation software, generating fictitious accounts, and providing mobile proxies is vast.

During recent years, the manipulation industry has become increasingly global and interconnected.

The openness of this industry is striking.

We developed this series of experiments and reports in support of the European Commission's Action Plan against Disinformation[15] and its original and now strengthened Code of Practice on Disinformation[16] to address the spread of online disinformation. The European Parliamentary Research Service referenced our 2019 and 2020 studies, echoing our earlier conclusion that platforms have not done enough to combat platform manipulation.[17]

INTRODUCTION TO THE EXPERIMENT

Our previous experiments assessing social media manipulation have furthered our understanding of the tools and techniques used to manipulate social media platforms. They have provided a framework for us to discuss specific issues with the social media companies to deepen our understanding of their abilities to counter platform manipulation.

This third iteration of our social media assessment has sought to enhance our methodology and deepen our cooperation with social media companies. During the year, we have had in-depth conversations with the assessed social media companies about our experiments, methods for countering abuse, as well as broader policy questions.

Like our previous experiments, the primary aim of this study is to test and assess the ability of social media companies to withstand manipulation from well-resourced commercial manipulation service providers. In this iteration of the experiment, we used three reliable Russian social media manipulation service providers to buy engagement on Facebook, Instagram, Twitter, YouTube, TikTok and—for the first time—VKontakte.

How is this relevant?

How is buying a thousand fake likes on a fake account relevant for the assessment of the overall ability of social media companies to counter manipulation? Manipulation is more likely to be reported if it engages with influential content relevant to current conversations. Indeed, a recurring theme in conversations with social media companies is that they prioritise moderating content that is likely to have a high impact and/or cause harm. They tell us that smaller manipulation efforts slide below their thresholds, that if the interventions were on a larger scale, they would be identified and penalised by their systems. Our interventions are designed to have low impact and be harmless for ethical reasons.

Yet, we argue that experiments of this type offer an effective way of assessing how platforms handle fake activity. The argument about scale only works in part—actors now wishing to achieve an effect spread the interventions over a longer period of time, a more diverse pool of fake accounts, on more individual messages. They have adapted, apparently successfully.

> The primary aim of this study is to test and assess the ability of social media companies to withstand manipulation from well-resourced commercial manipulation service providers.

We argue that attempts to separate and focus on activity by state actors versus commercial ones miss the bigger picture: manipulation relies on certain specific methods, infrastructure, and know-how. A company taking commercial clients today may be used by an actor seeking political ends tomorrow.

The manipulation we sample relies on the same technical abilities sophisticated state-backed actors would use to penetrate platform defences. Allowing the commercial industry to flourish has the added downside that there is a 'talent pool' from which state-actors can recruit.

The simple, cheap, commercial, and highly available manipulation relies on accounts that have a very specific footprint, which should be a low bar for the platforms and, therefore, constitute a relevant litmus test. If the platforms' algorithms do not spot this activity, they will be unlikely to detect a more determined actor automatically. The publicly available information we have from the recent Facebook leaks[18], past experiments conducted together with US senators and EU commissioners, as well as exclusive data analysis provided to us by Reset.tech, support our own analysis: suspected inauthentic engagement is not removed at a higher rate than what we report here.[19]

Our experiments show, at the very least, how fast, cheap, and effective it is to buy low-level commercial manipulation of social media platforms. That alone is worth tracking, assessing, comparing, and understanding as we try to further our understanding of inauthentic manipulation of the information space.

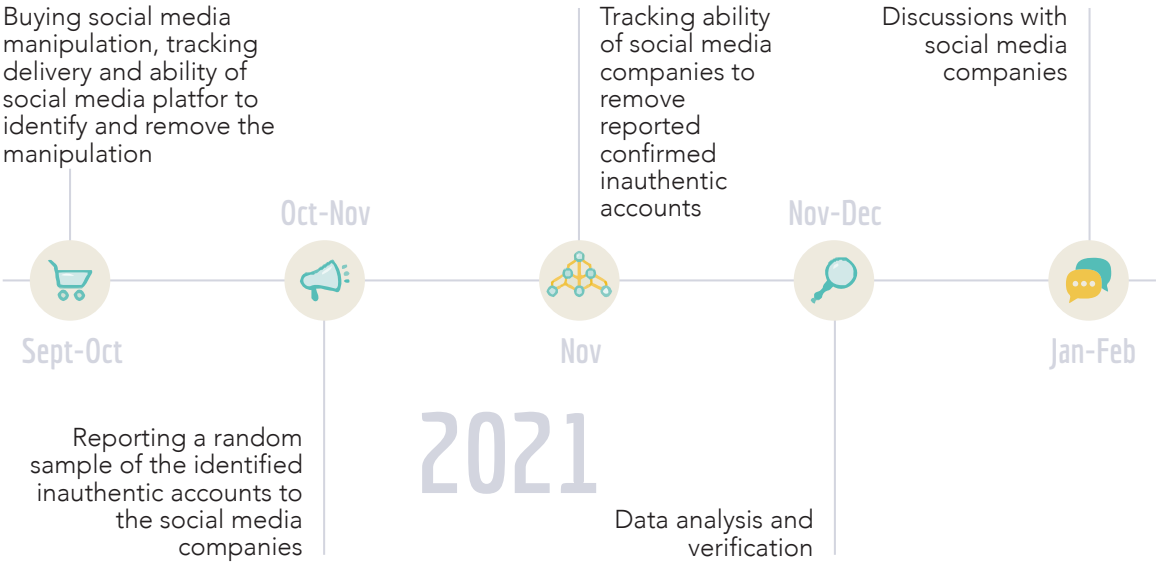
The Scale and Timeline of the Experiment

For the 2019 version of our experiment, we bought engagement on 105 different posts on Facebook, Instagram, Twitter, and YouTube using 16 different manipulation service providers. In 2020, we focused on three reliable providers and increased the quantities of engagement purchased. For our 2021 iteration of the experiment, we added another social media platform to our experiment.

In 2019, we spent €300 to buy 54,380 inauthentic engagements in the form of 3,530 comments, 25,750 likes, 20,000 views, and 5,100 followers, enabling us to identify 18,739 accounts being used for social media manipulation.

In 2020, we spent €300 and received 337,768 inauthentic engagements in the form of 1,150 comments, 9,690 likes, 323,202 views, and 3,726 shares on Facebook, Instagram, YouTube, Twitter, and TikTok, enabling us to identify 8,036 accounts being used for social media manipulation.

In 2021, we spent €279 and received 114,061 inauthentic engagements in the form of 1,385 comments, 13,859 likes, 93,009 views, and 5,808 shares on Facebook, Instagram, YouTube, Twitter, TikTok, and VKontakte, enabling us to identify 9,846 accounts being used for social media manipulation.



We conducted data collection during six weeks in September and October 2021. To assess the ability of the platforms to remove inauthentic engagement, we monitored our bought engagement from the moment of purchase to one month after it appeared online. We reported a sample of the inauthentic accounts identified to the social media companies and continued monitoring to measure the time it took for the platforms to react.

As part of the experiment, we recorded how quickly the manipulation service providers were able to deliver their services. We then collected data on how the six social media platforms responded to the manipulated content by periodically measuring whether it had been removed. The experiment was organised into the five steps visualised here:

The Ethics of the Experiment

Understanding how we can assess the abilities of social media companies to counter manipulation without causing harm or misinforming real world users has been our primary goal for the entirety of this research series. It would be possible to design an experiment that tests the ability of antagonists to manipulate political conversations by intervening in a real, ongoing political discussion, but such an experiment would risk undermining free speech.

We have set up these experiments to minimise risk and carefully monitor any inadvertent effects. To this end, we buy the fake engagement—views, likes, comments, and follows—using our own accounts created for this experiment. We continuously monitored our accounts to ensure there was no authentic human engagement with them. We also chose to engage with apolitical and trivial content. All purchased engagements were strictly designed and monitored to minimise impact on other online conversations.

Throughout the experiment, we did not observe any indication that our engagement had been noticed by authentic online users. Indeed, this was later confirmed to us by social media company representatives. For this reason, we concluded that we successfully managed to conduct the experiment without causing any harm to genuine online conversations.

Furthermore, we acted in the spirit of the social media companies' own white hat programmes; these programmes recognise the importance of external security researchers while emphasising the importance of protecting the privacy, integrity, and security of users. We spared no effort to avoid privacy violations and disruptions to real users; we did not access any real user data nor did we in any way attempt to exploit identified weaknesses for any reason other than testing purposes.[20]

Finally, we made every effort to minimise the amount of bought engagement, to avoid unnecessarily supporting manipulation service providers. We capped the amount spent at €279, which is consistent with the amount spent in previous reports.

> We successfully managed to conduct the experiment without causing any harm to genuine online conversations.

OUR ASSESSMENT CRITERIA FOR 2021

We assessed the performance of the six social media companies according to seven criteria measuring their ability to counter the malicious use of their services:

1. Blocking the creation of inauthentic accounts,
2. Removing inauthentic accounts,
3. Removing inauthentic activity,
4. Cost of services,
5. Speed and availability of manipulation,
6. Responsiveness
7. Transparency of actions.

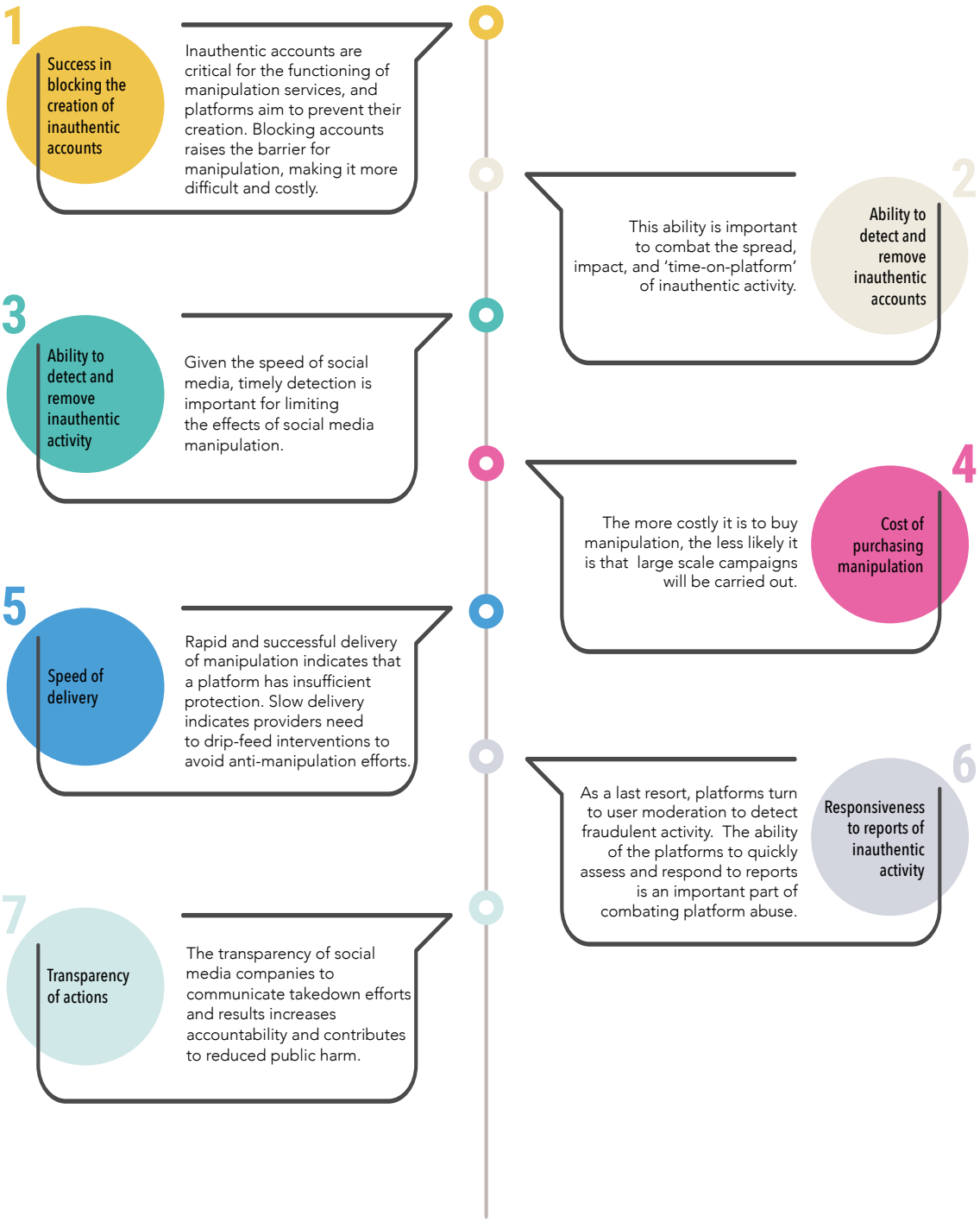
We have further developed and refined our criteria since our previous report.
We have not added any new assessment criteria, but we did slightly change how we measure our 'blocking the creation of inauthentic accounts' and 'transparency of actions' criteria.

In previous years, we had assessed the ability to block the creation of inauthentic accounts using a qualitative assessment of the measures put in place to prevent inauthentic account creation. However, this method does not allow for cross-platform comparison on a year-to-year basis. To enable a longer running comparison less sensitive to subjective assessment, we decided to use quantitative indicators instead—specifically, the price of purchasing inauthentic accounts.

We added an additional level to our 'transparency of actions' criterion. After we finalised our preliminary data, we sent a short summary and a sample of the manipulated content to each social media platform and asked for their comments on our experiment in general, as well as on the specific results provided to them. Their answers have been incorporated into our assessment.

These seven criteria can serve as general benchmarks for assessing the ability of platforms to counter social media manipulation.

> We have further developed and refined our criteria since our previous report.


1
Success in blocking the creation of inauthentic accounts
Inauthentic accounts are critical for the functioning of manipulation services, and platforms aim to prevent their creation. Blocking accounts raises the barrier for manipulation, making it more difficult and costly.
2
Ability to detect and remove inauthentic accounts
This ability is important to combat the spread, impact, and 'time-on-platform' of inauthentic activity.
3
Ability to detect and remove inauthentic activity
Given the speed of social media, timely detection is important for limiting the effects of social media manipulation.
4
Cost of purchasing manipulation
The more costly it is to buy manipulation, the less likely it is that large scale campaigns will be carried out.
5
Speed of delivery
Rapid and successful delivery of manipulation indicates that a platform has insufficient protection. Slow delivery indicates providers need to drip-feed interventions to avoid anti-manipulation efforts.
6
Responsiveness to reports of inauthentic activity
As a last resort, platforms turn to user moderation to detect fraudulent activity. The ability of the platforms to quickly assess and respond to reports is an important part of combating platform abuse.
7
Transparency of actions
The transparency of social media companies to communicate takedown efforts and results increases accountability and contributes to reduced public harm.

OUR ASSESSMENT FOR 2021

1. Blocking the Creation of Inauthentic Accounts

Blocking the creation of inauthentic accounts is perhaps the most important step social media platforms can take to prevent abuse of their platforms. The easier and faster it is to create fake accounts, the easier it will be to manipulate social media platforms. This is because access to the platform through fake accounts is the gateway to platform manipulation. Most, but not all, inauthentic engagement is delivered through the use of fake or inauthentic accounts.

We previously assessed the ability of social media platforms to counter the creation of fake accounts using a qualitative assessment based on an approach where we attempted to create multiple accounts and then tracked any pushback by the social media companies. During our experiment, however, we noticed both significant regional variation and significantly less pushback from the platforms, meaning that it was rather easy to create accounts on all the assessed platforms.

We now find that the detailed instructions from sellers of fake accounts regarding how to prevent blockage, such as preconfigured cookies to use with the account, dedicated proxies, specific IP locations that should be avoided, and other configuration descriptions, are primarily intended to prevent the accounts from being banned while in use rather than to stop the actual creation of a fake account. That said, several platforms have developed tools to counter the creation of fake accounts. For instance, Facebook sometimes requires users to record a video of their face to prove authenticity. Such features, however, are not uniformly enforced and can be evaded. While they add friction, they do not stop the creation of fake accounts.

Because it is easy to create fake accounts on all the platforms assessed, we decided to change our methodology in order to strengthen our assessment. As an alternative to qualitative assessment, we decided to use the price of inauthentic accounts as a quantitative indicator of how easy it is for the manipulation industry to systematically create fake accounts at scale. Our rationale is that the price of a fake account should correlate to the ease of account creation. The harder it is to create a fake account for manipulation, especially the more manual work is required, the more expensive it should be to purchase fake accounts.

Our collection and assessment of the price of fake accounts somewhat surprised us, as fake accounts on VKontakte are significantly more expensive than the other platforms assessed (see Figure 1). We had expected that Facebook (a platform with more visible protection against fake account creation)

would score highly; instead, they scored last. Among the four platforms with similar prices, Facebook accounts at €0.013 each are the cheapest, and Twitter accounts are the most expensive at €0.06. Over time, the prices are steady in relation to one another, but we see a slow, but steady, increase in price over time, signalling either increased friction from the social media platforms or simply inflation.

While there is movement in the right direction, it is still possible to buy many fake accounts at a low cost from the flourishing fake account industry. We see no indications that this is changing, and to date none of their actions have changed the basic functioning of the manipulation industry.

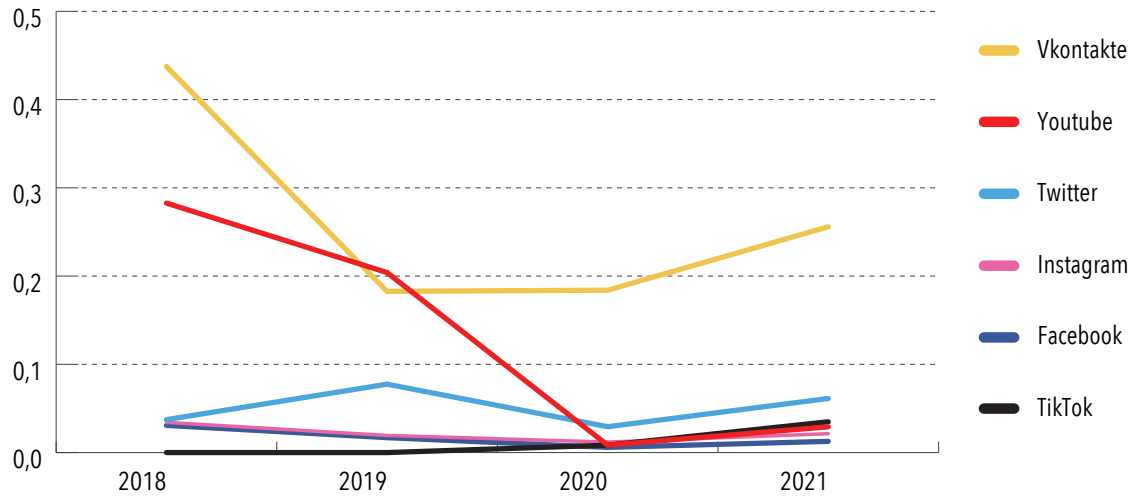


Figure 1. Average account prices for fake accounts, all platforms.

2. Removing Inauthentic Accounts

The longer bots and inauthentic accounts that are used for manipulation remain on a platform, the lower the cost for the manipulation service providers, as they don't have to spend time and money to replace blocked accounts. And by removing the fake accounts, together with associated activity, the impact of individual removals is amplified, potentially disrupting multiple larger operations. Because of this, removing the accounts is much more meaningful than removing individual posts.

From our discussions with several social media platforms, and indeed from the manipulation service providers themselves, we have learned that identification and removal of accounts engaging in manipulation are central components of the counter-manipulation strategies of all the social media platforms. Despite the millions of accounts reportedly

removed by social media companies each year for inauthentic behaviour, there is still a lack of insight and verification of the numbers reported by the companies. This means that there is insufficient public data to verify the accuracy of reported figures or, perhaps more importantly, what percentage of inauthentic accounts are removed. Our assessment offers some insight into how effectively active inauthentic accounts are removed on the various social media platforms that we have assessed.

During our past two assessments, roughly 20 per cent of the identified accounts were removed within our monitoring period. This time, 25 per cent of the identified accounts were removed on average. However, this improved result was largely due to including VKontakte in the experiment; VKontakte blocked or suspended 70 per cent of the identified accounts, a substantive difference compared to 14 per cent on the other platforms. VKontakte, on the other hand, does not remove the engagement of suspended or deleted accounts, meaning, for example, that comments left by a deleted account remain on VKontakte. VKontakte does, however, clearly indicate when a comment has been made by a removed or suspended user by changing the user's profile picture next to the comment. From our perspective, this design choice is unfortunate: while leaving the traces of sanctioned accounts visible may expose patterns of systematic abuse, users scrolling past on a timeline will only see the boosted numbers. Thus, the damage has been done, and it remains online when accounts are blocked or removed. Moderated content, however, is removed without a trace.

Comparing the half-life of accounts engaged in inauthentic behaviour (i.e., how long it takes for half of all identified accounts to be removed), only Twitter has consistently reduced the half-life of accounts engaged in inauthentic behaviour. VKontakte, on the other hand, quickly identifies and removes a majority of accounts engaged in inauthentic behaviour, albeit with little change after initial removal. On Facebook, Instagram, TikTok, and YouTube, very few accounts engaging in inauthentic engagement are identified and removed from the platforms.

Increased removal or suspension of accounts engaging in inauthentic engagement is desirable, as it increases the cost of platform manipulation. We believe that more platforms could do more in this field as shown by Twitter and VKontakte.

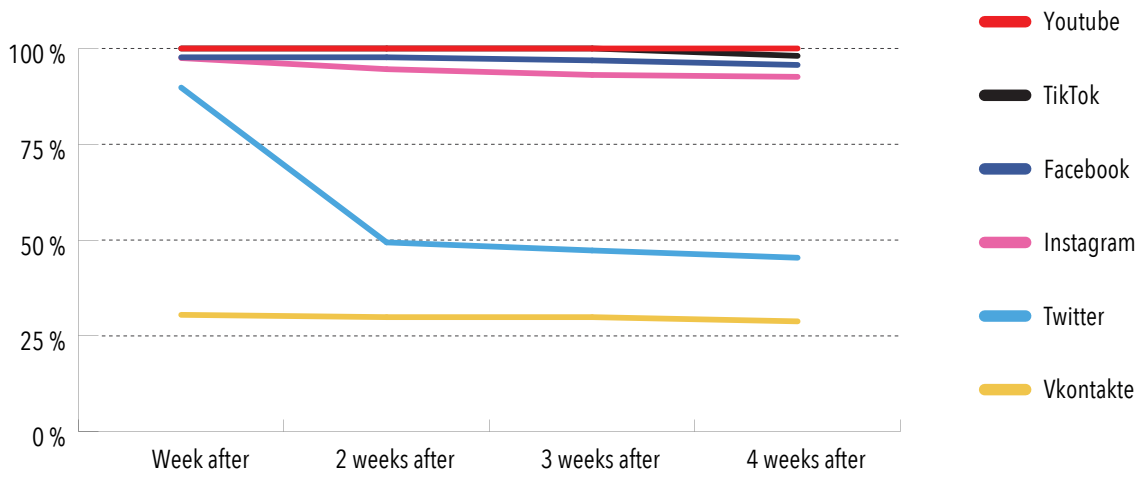


Figure 2: Comparison of half-life of inauthentic accounts by platform

	2019	2020	2021
Facebook	124	647	437
Twitter	68	41	25
Instagram	193	165	251
TikTok	*	*	999
YouTube	*	*	∞
VK	*	*	16

Table 1: Inauthentic account half-life by platform (days)
* Not available
∞ Infinite/no half-life due to lack of account removal

3. Removing Inauthentic Activity

Removing inauthentic activity is the process of identifying and removing fake engagement posted on the social media platform. The faster the inauthentic activity is removed, the smaller the effect the engagement will have on social media conversations, as fewer people will have had the chance to interact with the content.

In our previous two reports, we showed that social media companies struggled to automatically identify and remove fake activity, and that the vast majority of all the fake engagement was still online four weeks after delivery.

In the current experiment, TikTok performed significantly better than last time and improved from removing the least activity to removing the most activity, probably as a result of the platform's strengthened counter-abuse efforts during the past year. YouTube performed slightly better, while Twitter, Instagram, and Facebook performed worse. VKontakte performed marginally worse still.

Overall, 92 per cent of the inauthentic engagement remained active across all social media platforms four weeks later. For example, this means that if someone bought 1,000 fake likes on the account of their favourite, or least favourite, politician, brand, or restaurant, they could expect 920 likes to remain after a month. Fake views continue to be an especially potent problem as they are delivered instantaneously and seem to remain online indefinitely.

Social media companies continue to struggle to remove inauthentic engagement on their platform with marginal or no improvement from our previous report.

	2020	2021
Facebook	96.53%	98.52%
Twitter	74.23%	83.43%
Instagram	91.80%	96.01%
TikTok	99.69%	84.77%
VKontakte		99.96%
YouTube	97.17%	92.38%

Table 2: Percentage of activity remaining on the platforms after four weeks.

4. Cost of Services

The cost of manipulation is an indicator of how effectively social media platforms are combating manipulation. If accounts used to perform manipulation are removed, manipulation service providers have to spend time and money to replace them. When social media platforms redesign their service and render the scripts used to seed manipulation obsolete, developers have to update their scripts. These costs must ultimately be passed on to consumers. In discussion with several social media platforms, representatives noted that the cost of social media manipulation is an important indicator of their effectiveness in preventing platform manipulation.

For state-sponsored actors, the cost of these services is likely irrelevant—certainly the sums involved are trivial compared to those associated with kinetic military effects. But raising the cost is not so much about deterring state actors as it is to degrade the viability of the manipulation industry as a whole. State actors are more likely to maintain an in-house capability than rely on commercial services. However, in times of information contestation, activists supportive of individual politicians or governments may use commercial services in an attempt to dominate the online conversation. In this case, higher costs may well matter.

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 2021 and compared it to assessments of previous years as well as to historical data. The results do not indicate any significant shift in the cost of social media manipulation.

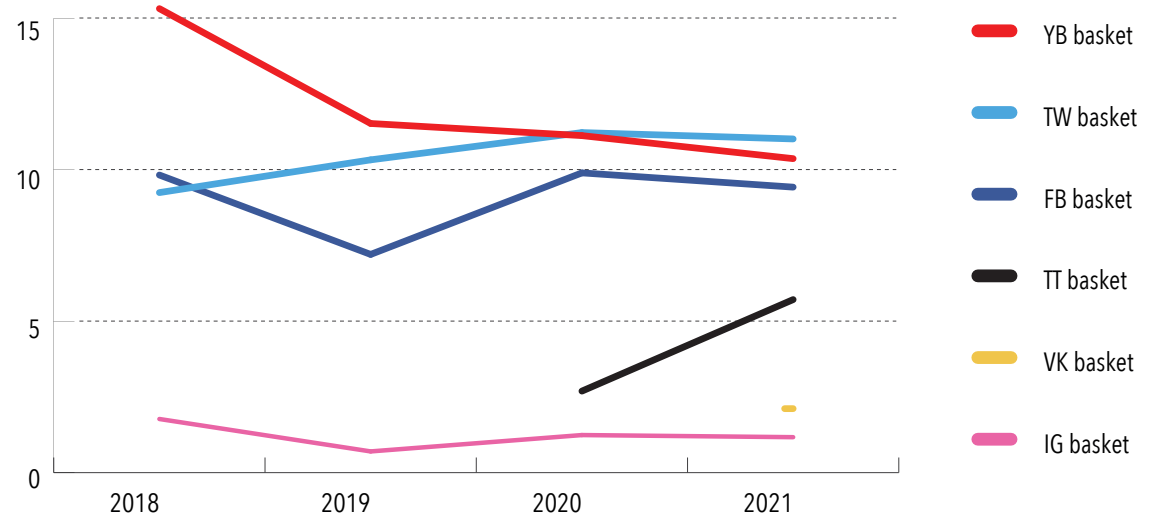


Figure 3: Price of a basket of social media manipulation.

Compared to 2020, we observe a slight decrease in the price of social media manipulation for all platforms other than TikTok, which is now twice as expensive to manipulate. We see this development more as a result of deliberate efforts made by TikTok, rather than the result of changing supply and demand dynamics surrounding a rapidly growing platform. Nonetheless, there may be a danger of backsliding once supply and demand forces settle down. Manipulation service providers continue to offer their services at roughly the same price as before, indicating that there hasn't been any significant change in the underlying conditions of the industry.

It remains less expensive to buy automatic manipulation, such as views and likes, while more labour-intensive manipulation, such as comments, are several times more expensive. It is interesting to note the significant price difference for followers between YouTube, Facebook, and Twitter on the one hand, and Instagram, TikTok, and VKontakte on the other, with the latter group being between five and fifteen times as expensive. For comments, Instagram and VKontakte stand out as especially cheap (see Figure 6).

For the third year in a row, Twitter is the most expensive platform to manipulate, and Instagram is the cheapest. This year's newcomer, VKontakte, scores second to last, being just marginally more expensive to manipulate than Instagram.

In 2021, on Instagram, 10 euros is enough to purchase almost 100,000 fake views, or 25,000 fake likes, 1,000 fake comments, or 15,000 fake followers.

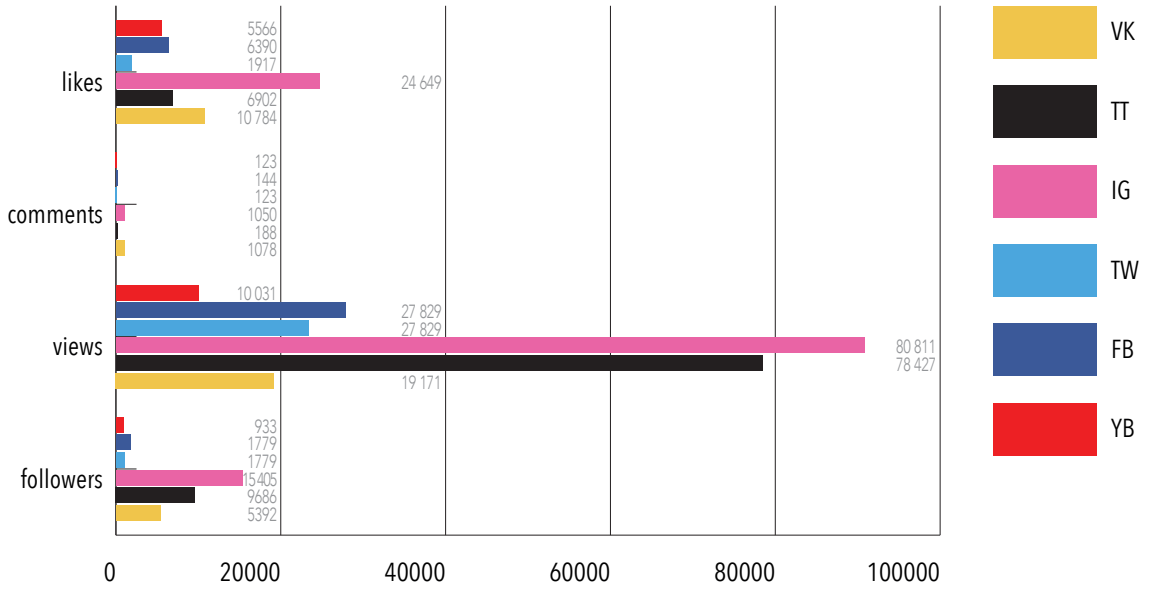


Figure 4: How much manipulation will 10 EUR buy you?

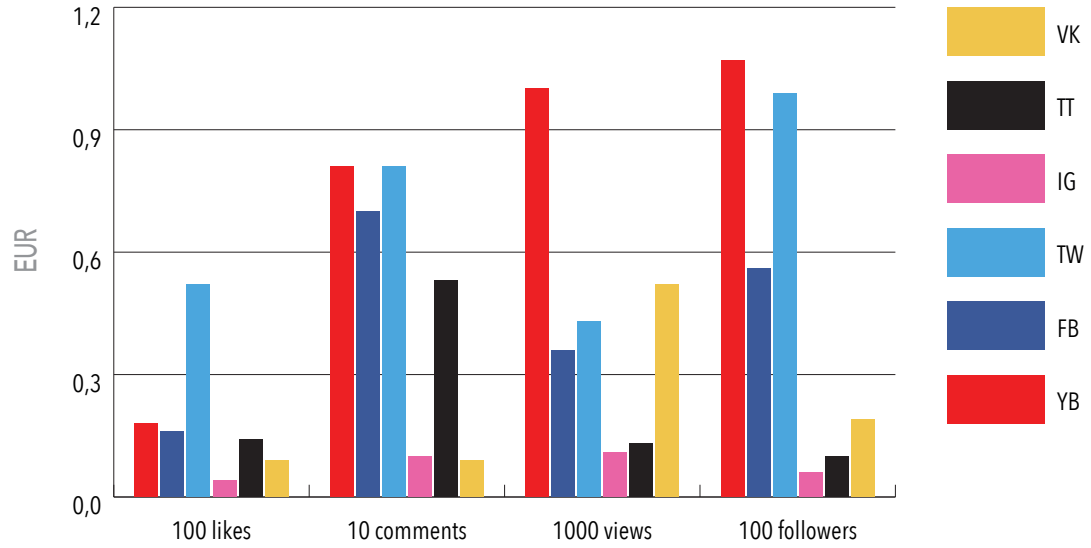


Figure 5: Median price for manipulation services across six providers

5. Speed and Availability of Manipulation

The speed and availability of manipulation is an important consideration for any antagonist seeking to influence the online environment. The faster you can deliver manipulation, the more likely it is that you can influence current events and discussions in a fast-paced digital environment. Slowing down the ability of manipulation providers to deliver fake engagement will reduce the impact and harm of the manipulation.

We note that, for platforms with ephemeral content, speed is especially important. This is particularly the case for Twitter, where material from yesterday is already buried under newer material. If these platforms can force manipulation providers to slow down, it is a big win. Conversely, much of the suggested content on YouTube is months or even years old. Here the manipulator has more time, which is to the platform's disadvantage. Thus, from the manipulators' perspective, if delays prevent effective manipulation on Twitter, they may be incentivised to manipulate YouTube instead. From this perspective, YouTube should be especially alert to the long-term challenge posed by fake interactions.

In 2020, we found that roughly 60 per cent of all manipulation bought across all platforms, excluding failed deliveries, were delivered within 24 hours. Manipulation on Twitter arrived most quickly, but was also removed the fastest. TikTok performed poorly, as manipulation was delivered almost instantaneously and remained over time.

In 2021, 77 per cent of all manipulation, across all platforms, was delivered within 24 hours, indicating that manipulation is delivered faster now than a year ago. TikTok performed significantly better than last time, with manipulation on TikTok delivered roughly

as fast as manipulation on Facebook and Twitter. YouTube performed significantly worse, with manipulation providers over-delivering: we received more than 100 per cent of the ordered volume within the first 24 hours. Manipulation on VKontakte was delivered at a slow but steady pace: By 72 hours, all manipulation had been delivered. TikTok, Twitter, and Facebook performed best, with less than 75 per cent of the manipulation delivered within 72 hours (see Figure 3).

On average, 20 per cent of all manipulation was delivered within an hour, and, after six hours, more than 30 per cent of the manipulation had been delivered on all the social media platforms. This indicates that it is still possible to manipulate current events using commercial service manipulation providers, and—unfortunately—manipulation is getting faster rather than slower.

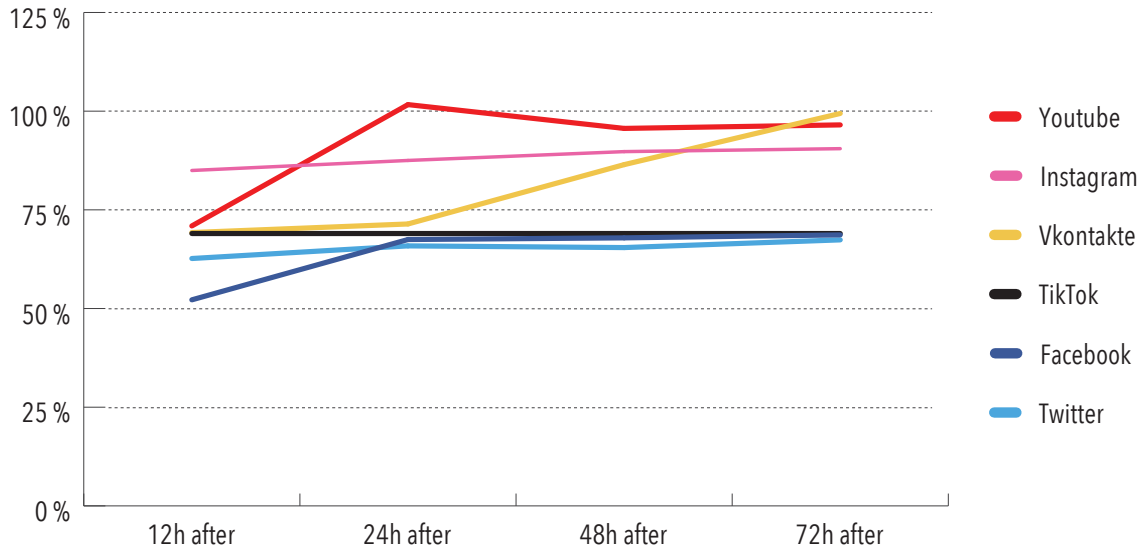


Figure 6: Comparing delivery speed of fake engagements—12, 24, 48, and 72 hours after purchase.

6. Responsiveness

To assess the responsiveness of the social media platforms, we reported between 50 and 150 random accounts from each platform that were identified as being used for social media manipulation. We then monitored how many of these accounts the platforms removed within five days.

In 2020, between zero and nine per cent of the reported accounts were removed, with several platforms removing no accounts at all. The current figures are very similar, with platforms variously removing between zero and ten per cent after five days (see Table 3).

In general, we observe that reporting an account does not lead to that account being blocked or suspended, even if the reported account is known to have engaged in inauthentic activity. We find that single reports of inauthentic activity do not trigger a review of the account and do not contribute to the removal of accounts engaged in inauthentic activity. This also holds for VKontakte, which seems to have a rather low threshold for suspending accounts suspected of engaging in inauthentic activity. Even after monitoring the accounts for fifteen days, the rate of blocking only changed in the case of Facebook, where we observe an increased rate—from ten per cent after five days, to twenty per cent after fifteen days.

We understand that social media companies might only take action once a minimum threshold number of users have flagged an account as harmful. The practice makes sense from the perspective of prioritising scarce moderation resources, but necessarily means that much potentially harmful content remains on the platform.

As in previous years, we conclude that reporting and moderation mechanisms must be improved so that a larger share of accounts flagged as inauthentic are acted upon, even if they are reported by a single user. It is problematic that inauthentic accounts, even when reported as such, typically escape penalty.

Facebook	90.00%
Instagram	99.00%
YouTube	100.00%
TikTok	96.08%
Twitter	98.00%
VKontakte	100.00%

Table 3: Share of accounts removed within five days of reporting.

7. Transparency of Actions

Transparency of actions enables researchers, as well as the public, to assess the efforts undertaken by social media companies to counter the manipulation of their services. The platforms that we have assessed offer varying degrees of transparency. For this iteration of the report, we combined an assessment of the platforms' public reporting content with the responses to a request for information sent to the social media companies. Our requests led to meetings with Twitter, TikTok, Meta (Facebook and Instagram), and Google (YouTube), as well as written analysis from Twitter and TikTok. VKontakte did not respond to our repeated requests.

We very much appreciated the opportunity to discuss our results with the platforms, as it improved our understanding of the choices the they make and why we got certain results. TikTok,[21] Twitter,[22] Facebook,[23] Google,[24] and VKontakte[25] all have pages dedicated to transparency reporting. Several companies have significantly improved their reporting and metrics since our last report. That said, there is more that platforms could do to increase transparency regarding inauthentic engagements to further enable independent assessment of the scale and nature of the problem. Currently, TikTok provides the most information about their work to enforce their community guidelines (for example, by reporting both the number prevent and the number of removed inauthentic accounts, videos, followers, and likes). This kind of in-depth reporting allows us to assess the scale of the problem in an unprecedented way.

In a remarkable disclosure, TikTok reported the prevention of 16.6 billion fake likes in a single quarter. Currently, no other platform reports the number of fake likes.

For the third quarter (Q3) of 2021, we assessed the amount of fake activity removed across the platforms which provide public data. In total, more than 22 billion spam, fake engagements, or fake accounts were prevented or removed during Q3 2021, reflecting the enormous scale of the problem.

Total interactions
19,3 Billion

Total interactions
2,6 Billion

Total removals
0,95 Billion

Twitter spam reports*
5 Million
* for Jan-Jun 2021

TikTok spam accounts prevented	226,557,055
TikTok spam videos removed	11,895,555
TikTok fake followers removed	231,150,220
TikTok fake follow requests prevented	2,078,453,724
TikTok fake likes removed	203,708,379
TikTok fake likes prevented	16,594,976,202
TikTok fake accounts removed	1,728,621
TikTok total interactions	19.3 Billion
Facebook fake accounts actioned	1,800,000,000
Facebook spam accounts actioned	777,000,000
Facebook total interactions	2.6 Billion
YouTube spam channels removed	3,457,303
YouTube spam comments removed	950,872,903
YouTube spam videos removed: 339,763	339,763
YouTube total removals	0.95 Billion
Twitter spam reports	5,144,026 (*for Jan-Jun 2021)

Social media companies use differing terminology for reporting how they enforce their community guidelines, often mixing inauthentic engagement with other kinds of spam, making it difficult, or impossible, to separate the different kinds of fake engagements. TikTok details the kind of inauthentic engagement both prevented and removed, thereby providing significant insight. To further improve transparency, we would also recommend a relative figure (such as the number of removed or prevented likes in relation to the total number of likes on the platform). More details about other forms of manipulation, such as fake views, would also be helpful. But the single biggest improvement would be a detailed qualitative quarterly or annual assessment by the platforms themselves of the manipulation that was stopped and removed, as well as what content was targeted.

The difference in the numbers reported by the platforms is reflects varying terminology, definitions, and willingness to release data. There is nothing to suggest that TikTok has a larger problem with inauthentic behaviour than Instagram simply because the self-reported figures differ.

In terms of reporting on inauthentic behaviour, Meta (Facebook and Instagram)

delivers industry-leading qualitative reports. Honourable mentions also go to Twitter and Google. Twitter regularly provides useful data sets to the research community. VKontakte and TikTok do not provide such reports or data sets on coordinated inauthentic behaviour.

Our conclusion from previous reports holds true: much more can still be done in terms of contextualising information, conducting thorough audits and publishing the results, developing disclosure frameworks and collaborative transparency, and formulating best practices jointly with other platforms, etc. Efforts to provide researchers with data access have improved, with Twitter in particular and Facebook, to a lesser extent, offering access to researchers and releasing specific data sets of inauthentic coordinated behaviour.

It remains difficult to independently assess how the threat landscape is developing and how well counter efforts are working based on the information currently provided by the platforms. Overall, much more can be done to increase transparency to enable researchers and users to independently assess and compare platform performance.

ASSESSMENT OF EACH PLATFORM'S RELATIVE STRENGTHS AND WEAKNESSES

The performance of each social media platform is assessed using the criteria introduced above and qualitative ratings, based on a joint assessment by the research team. These ratings are then used to assess the relative performance of each of the social media platforms.

Facebook

Previously, we had always assessed that Facebook was the most effective platform when it came to blocking the creation of fake accounts. This time, however, Facebook performed significantly worse under our updated methodology for assessing the ability of platforms to prevent fake account creation. This change in methodology was partially initiated by a discovery that Facebook does not deploy its advanced counter-fake technology uniformly in all its markets, meaning that it is possible to evade them. Indeed, we had the impression that it was easier to create accounts in 2021 than in previous years, when we were regularly challenged to perform various tasks such as completing a captcha or submitting a selfie.

Rating Facebook's ability to counter fake account creation on the basis of the price of fake accounts gave a very different result from our previous qualitative assessment. Rather than scoring first, Facebook now scores last since Facebook accounts are the cheapest accounts to buy.

In 2020, we concluded that Facebook was the only platform with an increase in the half-life of active inauthentic accounts. The latest results indicate a half-life of 437 days, which is an improvement compared to 2020, but worse than in 2019 and considerably worse than Twitter, which has managed to lower the half-life of fake accounts for three consecutive years—from 68 days to 25 days.

Compared to our previous report, Facebook's record for removing inauthentic content has worsened, as the percentage of removed activity fell from 3.5 per cent to 1.5 per cent. We also did not observe the gradual removal of inauthentic content we had observed previously, which together meant that Facebook performed substantially worse in relation to the other platforms this time.

Assessing speed of delivery, Facebook performed well, capping deliveries under 70 per cent for the first 72 hours. Facebook deliveries were also the slowest after 24 hours (which is good), in line with the results observed in our previous report.

The price of Facebook manipulation remained stable; it remained in third position behind Twitter and YouTube. Facebook did, however, maintain its position

as the platform which removed the highest percentage of reported accounts and slightly increased from 9 to 10 per cent. That ratio, however, is still too low to effectively counter manipulation attempts.

In terms of platform transparency, Facebook continues to perform well and deliver the best coordinated inauthentic behaviour assessments in the industry. In relation to transparency regarding inauthentic engagement, however, Facebook has been leap-frogged by TikTok's new and more granular transparency initiative. We remain unsure regarding important details on how Facebook compiles its statistics and would welcome independent assessment and auditing of the data and reports published by the platform (something which holds true for all the platforms).

In total, Facebook performed worse than in 2020 in four out of seven categories, scoring a total of 25 points and thereby now sharing the runner-up position with TikTok.

Instagram

Even though Instagram is owned by Meta, it is far less effective at countering platform abuse. In fact, Instagram is outperformed by Facebook in all but one assessed category. Manipulating Instagram still remains cheap and effective. With TikTok having made significant improvements since 2020, Instagram once again emerges as the least effective platform at countering inauthentic engagement. Even VKontakte, included for the first time, ranked higher than Instagram in many categories.

TikTok has demonstrated that it is possible to make rapid progress in combating inauthentic activity. Instagram's lack of improvement can only be interpreted as an intentional decision not to prioritise this work. Meta might consider sharing knowledge and expertise from Facebook with Instagram to build platform-specific defences that are uniformly effective across their services. The fact that Meta now offers a simple toggle between Facebook and Instagram in Meta's transparency reporting, and that the Instagram view often shows missing data, is hopefully a sign that work is also in progress here. We hope to see improvements from Meta and Instagram during the coming year.

The single category where Instagram outperformed Facebook is account removal. Instagram removed 7.3 per cent after four weeks. For content removal, though, Instagram performed worse this year and is now on par with Facebook's numbers.

Instagram's failure to counter manipulation is still most evident in the cost of manipulation, where Instagram continues to be the least expensive platform to manipulate, slightly cheaper than in 2020. We also realised this year that the transparency figures reported by Meta do not include any figures from Instagram, prompting us to lower their transparency rating.

In assessing Instagram's performance in comparison to 2020, we observed a slightly deterioration. We therefore continue to conclude that Instagram remains very easy to manipulate, and Facebook should prioritise sharing its expertise with Instagram to strengthen their counter-manipulation work.

Twitter

For the third year in a row, Twitter is the most effective platform overall at countering manipulation. That said, in 2021, Twitter performed worse in several categories and lost a total of three points in our combined ranking.

Twitter removed almost ten per cent less inauthentic engagement in 2021 as compared to 2020. While Twitter still removed more than the other platforms, this is a significant drop in the platform's performance.

The cost of manipulating Twitter remained stable from 2020 to 2021. More worrying is that the amount of manipulation delivered within 72 hours increased by more than 20 per cent, from 43 to 69 per cent, positioning Twitter on a level with TikTok and Facebook in this category. Facebook performs slightly better than both Twitter and TikTok by virtue of having less manipulation delivered within the first 12 hours.

Twitter provided in-depth feedback to our request for information based on the preliminary data that we provided.

Twitter determined that 90 per cent of accounts that interacted with our content had since been removed, due to a combination of automated enforcement and actions taken by Twitter's investigators. Twitter also noted that the content we had bought manipulation on got little, or no, authentic engagement, indicating that the effect of the manipulation was limited.

Twitter further detailed that, in 2021, they introduced new measures to protect against spam– both automated and human-coordinated. For example, to

> We, therefore, continue to conclude that Instagram remains very easy to manipulate.

> Twitter maintains its position as the most effective platform at countering platform manipulation.

address so-called "copypasta" campaigns, Twitter began filtering duplicate text to the "show more" replies in conversations. As an extension of this work, spammy, duplicative content is, according to Twitter, no longer visible in the platform's "Top" search. Twitter also noted that they continue to provide access to comprehensive data sets about state-linked manipulation operations to researchers and that they have now expanded access to data beyond information operations to include the Twitter Moderation Research Consortium (TMRC)—a global group of experts from across academia, civil society, NGOs, and journalism, studying platform governance issues. Members will conduct their own independent research on said data, publishing research insights, aimed at deepening understanding of these challenges. We look forward to seeing how this develops in the coming year.

Twitter maintains its position as the most effective platform at countering platform manipulation. Their performance reduction, however, in our assessment, is somewhat worrying and warrants review by Twitter. Furthermore, Twitter could still do more to block the creation of fake accounts, to increase their ability to remove reported inauthentic accounts, and to further improve the granularity of their transparency reporting related to inauthentic activity.

YouTube

Over the past three years, YouTube has steadily been gaining points in our assessment, but they continue to struggle to introduce sufficient friction against inauthentic manipulation.

In 2021, the cost of manipulation on YouTube stayed the same, and they removed slightly more inauthentic engagement. The speed of delivery, however, increased significantly. Faster manipulation means antagonists have a greater chance of achieving real impact. After 24 hours, more than 100 per cent of the bought engagement had been delivered, in comparison to 61 per cent in the previous iteration of this experiment.

YouTube still provides less by way of facts, statistics, and takedown reports than Facebook, Twitter, and TikTok. However, Google's Threat Analysis Team bulletins have improved and now provide valuable additional details.

It should be embarrassing that it is still easy to find YouTube tutorials explaining how to buy manipulation on YouTube and many other platforms. Furthermore, manipulation service providers continue to advertise their services successfully through Google Ads—YouTube's parent company. While we note that ads for manipulation seem to have been banned in English, searches in other European languages still generate ads for manipulation services. Google needs to deal with this problem as it clearly undermines their own platform as well as other social media platforms.

TikTok

TikTok has made remarkable improvements since our last assessment, when we identified significant flaws on their part. Since our last report, we have had several conversations with TikTok and discussed ways to improve their ability to counter inauthentic manipulation. TikTok also provided specific feedback to our preliminary findings. It is clear that their serious effort has paid off, given that TikTok has gone from being last to the runner-up in our assessment in just over a year. It further strengthens our core argument that platform protection is related to the effort spent by the social media platforms to counter abuse.

We commend TikTok's transparency efforts, which are cutting-edge in this specific field (inauthentic activity), although they still lack insightful reports regarding coordinated inauthentic behaviour. We also note that TikTok is using notifications to inform its users when they have found inauthentic engagement and intend to remove it. This is an encouraging innovation and something we have been advocating for some time.

TikTok's main improvements have been in removing inauthentic engagements, indirectly increasing the price of inauthentic engagements, and reducing the speed of delivery. TikTok only made marginal improvements when it came to removing inauthentic accounts and responding to user reports; here there is still significant room for improvement. Another area where TikTok can improve is through offering data and tools for external researchers. Currently, it is hard to search the platform or collect the data collection necessary to identify examples of coordinated inauthentic behaviour.

VKontakte

VKontakte surprised us with several novel approaches to countering inauthentic engagement. The most visible difference is that they clearly identify whether an account has been suspended or removed, and, even while a comment from a suspended or removed user remains, it is clearly visible that the comment is made by such a user. VKontakte also seems to have a much lower threshold for blocking accounts and forcing users through a captcha and email verification process.

And, lastly, VKontakte is the only platform which mandates a mobile phone number for registration, significantly raising the costs for registering fake accounts, something that is also reflected in the price of a fake account. In total, VKontakte has a different approach to addressing the problem.

How well does this approach work? We assess that it has strengths and weaknesses. It does raise the bar for creating a fake account, but not enough, given that there are multiple services for buying temporary phone numbers for registration purposes. Their approach also removed a much larger number of fake accounts than the other platforms, but not enough to effectively hinder manipulation, which is reflected in the low cost of manipulating VKontakte. The low cost might be a reflection of VKontakte's inability to remove inauthentic engagement over time. VKontakte, much like YouTube, also failed to act on any of our user reports of inauthentic accounts.
VKontakte has a transparency page like the other platforms but it is not up-to-date (with its last entries being from 2020), and it does not provide any specific data relating to inauthentic engagement. VKontakte is the only platform that publicly lists their platform integrity team's contact details. However, as they never replied to us, it's difficult to assess the usefulness of this transparency. We also note that by all accounts VKontakte has a cosy relationship with the Russian security services, an altogether more negative form of transparency.

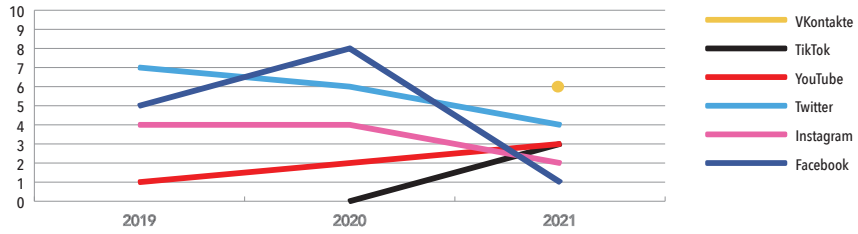
It is interesting that VKontakte has found a way to identify and remove inauthentic accounts much more quickly and accurately than the other social media platforms that were assessed, but, regrettably, for VKontakte that does not translate into better protections overall. In the end, they performed slightly better than Instagram and YouTube but are significantly behind TikTok, Facebook, and Twitter.

Relative Performance

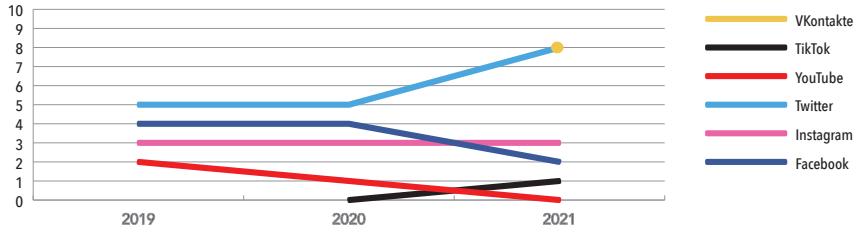
The most important insight from our study continues to be that there is a significant difference among platforms in their ability to counter manipulation of their services. This is an important insight for policy makers, regulators, and the companies themselves.

Twitter remains the industry leader in 2021, with TikTok and Facebook close behind.

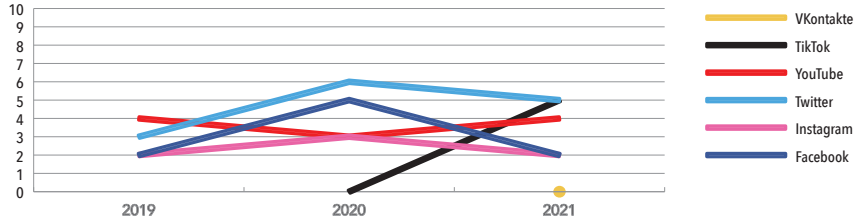
1. BLOCKING ACCOUNT CREATION



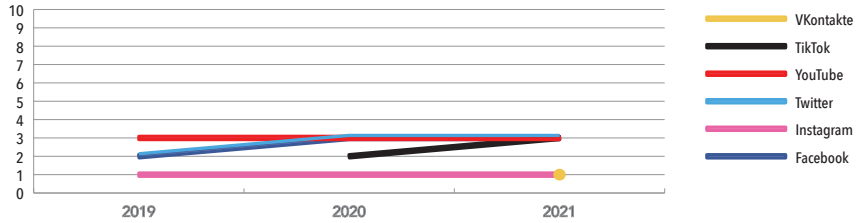
2. REMOVING ACCOUNTS



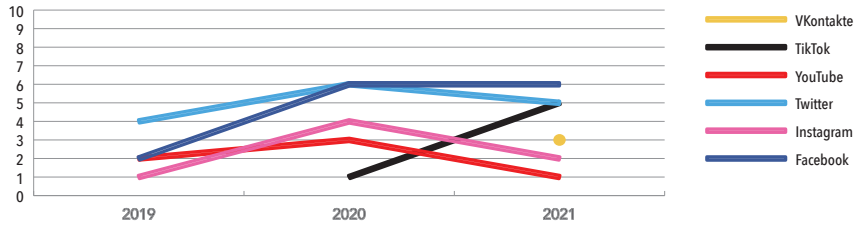
3. REMOVING ACTIVITY



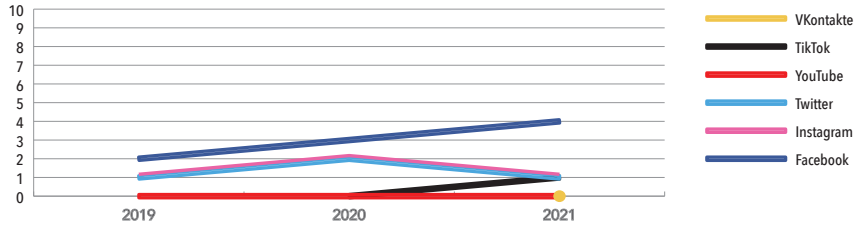
4. MANIPULATION COSTS



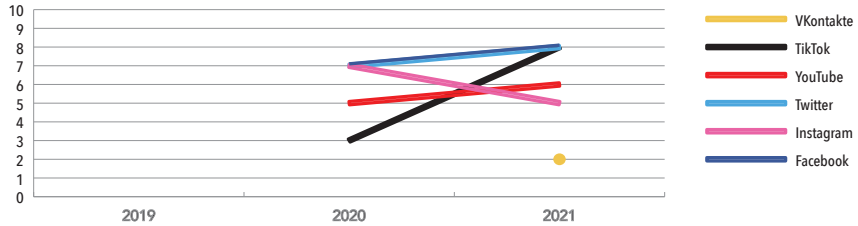
5. SPEED OF DELIVERY



6. RESPONSIVENESS



7. TRANSPARENCY OF EFFORTS



TikTok improved significantly during 2021, and the newcomer, VKontakte, performed well thanks to its ability to block fake account creation and remove inauthentic accounts.

Despite notable improvements by some, none of the six platforms we studied are doing enough to prevent manipulation of their services. The manipulation service providers are still winning the digital arms race. It is still easy, cheap, and quick to manipulate social media platforms.

What Other Content Were the Bots Targeting?

In this section, we present an overview of other content that the accounts used to deliver manipulation services to us interacted with. In some cases, it was hard to get access to sufficient data to identify the accounts; for instance, on YouTube and TikTok, we could only see accounts used to place comments. In some cases, when we could identify an account as being responsible for specific inauthentic manipulation efforts, we were unable to see the other activities of that account.

For this iteration of the report, we were able to confidently identify and track the activity of some accounts used by the manipulation providers to boost our posts (see Figure 7).

Some themes emerged as common across all of the platforms. In every case, we identified that accounts had been used to boost the visibility of online influencers and celebrities, online games, cryptocurrencies, and various online financial services. On VKontakte and Twitter, the accounts had interacted with pornographic content. On Facebook, accounts promoted a company claiming to provide 'client data lists'.

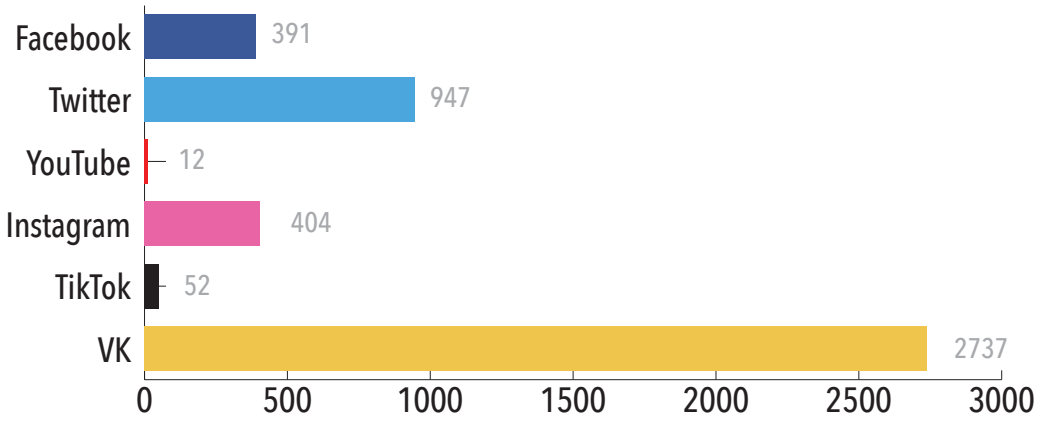


Figure 7: Number of inauthentic accounts tracked

Sadly, as in the 2020 version of the report, we observed attempts to spread disinformation surrounding the COVID-19 pandemic and vaccines. This time, we found such material boosted on Facebook, Twitter, and VKontakte.

We identified a lot of material promoted by the tracked accounts relating to the 2021 Russian Duma elections and various Russian politicians at various levels. The most visible examples related to candidates of the Liberal Democratic Party of Russia (LDPR), which we identified on Twitter, Instagram, and YouTube. Additionally, we identified promoted material in favour of the ruling United Russia party, the Communists (KPRF), and the New People Party, which many consider a spoiler party.[26] We also found accounts pushing an ad attacking the Communists.

On both Twitter and VKontakte, the tracked accounts promoted material posted by a recently sanctioned Russian TV personality. Other noteworthy examples from Twitter include a well-known pro-Kremlin troll account, anti-Navalny content posted by a since-suspended pro-Kremlin blogger and political entrepreneur. On Facebook, we identified accounts posting in support of a Russian foreign policy think tank headed by a United Russia Duma deputy.

We also identified a long list of international actors apparently benefiting from the publicity generated by the same manipulation providers.

- Head of a Saudi government agency (Twitter).
- Advisor to the Saudi Royal Court (Twitter).
- A Lebanese journalist (Twitter).
- A number of accounts associated with the National Council of Resistance of Iran (Twitter).
- A prominent Venezuelan politician (Twitter).
- Various accounts from public figures in Indonesia, Egypt, Brazil (Twitter).
- A member of the Verkhovna Rada of Ukraine (TikTok).
- A Turkish mayor (YouTube).
- A Ukrainian deputy mayor (YouTube).
- An Indian computer science professor (YouTube).
- A member of the Hong Kong Legislative Council (YouTube).
- A Nigerian firebrand pastor (YouTube).
- One dentist, a health guru, a number of medical clinics (YouTube).
- The director of a Ukrainian pet food business (Facebook).
- A former Ukrainian president (Facebook).
- A former Ukrainian government minister (Facebook).
- A member of the Italian senate (Facebook).

These examples provide insight into the diverse cast of character types who gained an inauthentic increase in their visibility online. In none of the cases above is there any firm evidence that the organisations or individuals apparently benefitting from the fake interactions were in any way involved in their purchase. It is possible to artificially boost a political opponent as part of a smear-campaign, but we didn't see any suggestion of such intentions here.

Our conclusion from previous reports stands: manipulation services are still being used primarily for commercial purposes, but political actors are making regular forays into manipulating public discourse.

CONCLUSIONS

This study demonstrates that social media companies struggle to identify and remove inauthentic activity. The vast majority of all the fake engagement delivered was still available four weeks after delivery. Even if there is no dramatic change since our 2020 study, most of our indicators now point in the wrong direction. Social media manipulation is, on average, faster and cheaper than it was one year ago.

Overall, 92 per cent of the inauthentic engagement remained active across all social media platforms after four weeks. In practice, this means if someone bought 1,000 fake likes on their favourite, or least favourite, politician, brand, or restaurant, they could expect 920 likes to remain after a month. They could also expect 77 per cent of all manipulation, across all platforms, to be delivered within 24 hours, or 20 per cent within an hour.

Our main insight from this and previous assessments is that platforms are not equally bad. This year it is once again clear that investment, resources, and determination make a significant difference in the ability of social media companies to counter manipulation. TikTok's sharp improvement is a stark reminder that social media companies can do more, and they can do better at countering platform manipulation. Our assessments should also be seen as a continuous reminder for policymakers that we need more transparency in this field to be able to understand the true scope of platform manipulation. Only then will we be able to develop meaningful joint counteractions against the manipulation service providers.

Many of the challenges the companies face could be addressed more effectively if they improved communications, established forums, and chose to work jointly to combat the problem at hand. The need to work together has never been more evident—not only for social media companies but for our society as a whole. Effective counter efforts can likely only be achieved if telecom companies, online payment providers, web hosting services, search engines, and online advertising companies all come together to combat the digital manipulation industry.
It is still far too easy to manipulate social media platforms, and the overall improvement witnessed before has stagnated. Even if individual companies are improving, our general assessment is that social media manipulation continues to be cheap and effective. Much more needs to be done to counter the manipulation industry and those who seek to undermine social media.

Even if it isn't the focus of this study, it remains evident from the published takedowns regarding state-driven social media manipulation that antagonists continue to exploit social media loopholes to manipulate

public discussions. We assume that the skill sets developed by the manipulation industry contribute to the skill set and technical know-how available to state actors as well.

Policy Recommendations

The policy recommendations we presented in our initial report remain in force. The developments we have observed over the last year strengthen our conviction that our original recommendations are important and remain much-needed.

1. Increase transparency and develop new safety standards

We have had many discussions with social media companies regarding the accuracy of our assessments, discussion we deeply appreciate as they greatly enhance our understanding of the problem. One of their main counter-arguments is that we are unable to assess their ability to identify and take down large scale operations conducted by sophisticated actors—and perhaps even more important—to limit the reach and visibility of manipulation. We fully agree with them and we acknowledge the limitations of our experiments—but we cannot conduct large-scale manipulations to test their arguments.

We are also much more concerned about the platforms' ability to detect and counter large-scale, state-backed information operations than relatively low-level spam. But our perspective differs in that we see the techniques, tools, and sometimes even vendors of commercial manipulation as part of the toolkit used by more sophisticated actors as well. We know that the notorious Internet Research Agency maintained commercial clients alongside its more famous operations on the Kremlin's behalf. It is a similar story in the world of cyber exploits, where hackers are pressured or otherwise incentivised to support the work of various civilian and military agencies. If the social media companies want their algorithms to detect and block sophisticated examples of social media manipulation, catching primitive examples of it would seem a natural first step.

We require better transparency to assess how effectively social media companies counter manipulation. In particular, more detailed information is needed regarding the scope and effect of manipulation as well as the 'actors, vectors, targets, content, delivery mechanisms, and propagation patterns of messages intended to manipulate public opinion'.[27] In essence, it is important to know who is trying to manipulate social media platforms and to what effect. The current transparency reports tell us that billions of fake accounts are removed from the platforms every year, but we don't know what these accounts sought to achieve, or indeed what effects they delivered before being identified. To assess their impact on social media conversations, business, online advertising, and ultimately our democratic discourse, more transparency is needed.

Furthermore, we need a common safety standard and standardised reporting that allows watchdog agencies to compare reports from the different social media companies. Tech companies also need to be encouraged, or obligated, to share technical data and know-how that would enable joint development of best practices and optimise capabilities for tracking and removing antagonists across platforms.

Finally, a system of independent auditing should be considered in order to build and maintain trust in the reports from the social media companies.

2. Establish independent and well-resourced oversight

Independent oversight would help provide the insight needed to better assess the progress of the social media companies in countering inauthentic activity on their platforms. Given the wide variation in the ability of social media platforms to counter manipulation, it is becoming ever clearer that impartial and independent assessment of the effectiveness of social media companies' attempts to counter platform manipulation is a necessity.

3. Deter social media manipulation

While we have focused on the ability of social media companies to protect their platforms, it is also important that we turn our attention to the industry that profits from developing the tools and methods that enable this interference. Lawmakers should seek to stop the social media manipulation industry by making it illegal to buy and sell social media manipulation. The ongoing practice of widespread and relatively risk-free social media manipulation needs to stop.

4. Social media platforms need to do more to counter abuse of their services

Even though we have observed important improvements by some social media companies over the past years, it is important that we continue to pressure them to do more to counter platform manipulation as we have a long way to go to make manipulation slow, expensive, and ineffective. Manipulation service providers continue to advertise and promote their services on the very platforms that they seek to undermine. It is striking that some social media companies are unable to prevent the manipulation service providers from using their own platforms to market services designed to undermine platform integrity. It may well be that the incentives for platforms to tackle the problem are insufficiently strong—after all, fake clicks also generate advertising revenue.

It is especially embarrassing that Google's search engine continues to allow ads from the social media manipulation industry. Though not included in this study, it is

worth noting that the same applies to Microsoft's search engine, Bing. These companies continue to profit financially from social media manipulation, despite the fact that we have highlighted this problem for several years.

5. A whole-of-industry solution is needed

Social media companies will not be able to combat social media manipulation without a whole-of-industry solution to the problem. Payment providers such as PayPal, Visa, and Mastercard should stop payments to the manipulation industry. Advertisers need to further push social media companies to counter abuse of their platforms and to sanction influencers who use social media manipulation to defraud the ad industry. By undermining the commercial manipulation of social media platforms, we will also make it more difficult for politically motivated actors to use the same technology and tools to manipulate political conversations.

Implications for NATO

Social media manipulation continues to be a potential challenge for NATO; it is a potent tool for malicious actors seeking to undermine the interests of the Alliance. For years, antagonists have continued to improve their skills and ability to compromise and exploit social media conversations. Disclosures by social media companies continue to underscore the intensity and determination of foreign states and other antagonists to undermine the interests of the Alliance.

Our assessment shows that commercial manipulation services are cheap, fast, and effective. We assume that the know-how, techniques, and services readily available by Russian commercial manipulation services can be used by state-actors to threaten the interests of the Alliance.

As the defences of the social media companies are still inadequate, we must continue to expect that antagonists will be able to exploit social media for malicious purposes during times of peace, of crisis, and of war. Therefore, the Alliance must continue to develop and refine its strategies and its ability to communicate in a highly contested information environment.

NATO member states and related international bodies, such as the EU Commission, must also continue to press social media companies for disclosures outlining their ability and effectiveness in combating commercial social media manipulation, as well as state-sanctioned hostile social media manipulation.

ENDNOTES

1 CNBC, 'Facebook, Twitter remove disinformation accounts targeting Ukrainians'. [Accessed 14 March 2022].

2 Meta, 'Updates on Our Security Work in Ukraine'. [Accessed 14 March 2022]

3 Bellingcat, 'Exploiting Cadavers 'and 'Faked IEDs': Experts Debunk Staged Pre-War 'Provocation' in the Donbas'. [Accessed 14 March 2022].

4 EUvsDisinfo, 'The Kremlin's playbook: Fabricating pretext to invade Ukraine - more myths. [Accessed 14 March 2022].

5 The Guardian, "Dumb and lazy': the flawed films of Ukrainian 'attacks' made by Russia's 'fake factory' [Accessed 18 March 2022].

6 Atlantic Council. 'Why Vladimir Putin is losing the information war to Ukraine' [Accessed 14 March 2022].

7 The Guardian, 'Game of Whac-a-Mole': why Russian disinformation is still running amok on social media, December 2018.

8 NATO StratCom CoE, The Black Market for Social Media Manipulation, December 2018.

9 Sebastian Bay and Rolf Fredheim, 'How Social Media Companies are Failing to Combat Inauthentic Behaviour Online', 2019.

10 Sebastian Bay, Anton Dek, Iryna Dek, and Rolf Fredheim, 'Social Media Manipulation Report 2020', 2020.

11 Jens Mattke, Christian Maier, Lea Reis, and Tim Weitzel, 'Herd Behavior in Social Media: The Role of Facebook Likes, Strength of Ties, and Expertise', Information & Management 57, № 8 (December 2020).

12 Vice News, 'Russian TikTok Influencers Are Being Paid to Spread Kremlin Propaganda [Accessed 16 March 2022].

13 Johan Lindquist, "Good Enough Imposters: The Market for Instagram Followers in Indonesia and Beyond", i The Imposter as Social Theory: Thinking with Gatecrashers, Cheats and Charlatans (Bristol: Bristol University press, 2021).

14 NATO StratCom CoE, 'Robotrolling'. [Accessed 10 February 2022].

15 European Commission, 'Action Plan on Disinformation'. [Accessed 10 February 2022].

16 European Commission, 'Code of Practice on Disinformation'. [Accessed 10 February 2022].

17 European Parliamentary Research Service, 'Key risks posed by social media to democracy'. [Accessed 10 February 2022].

18 Wallstreet Journal, The facebook files. [Accessed 16 February 2022].

19 We were given access to the primary data used for the assessments referenced in Reset.tech and Hate Aid, '210831_Reset_Facebook_Bundestagswahl_EN.pdf'. [Accessed 10 February 2022].

20 See, for example: Facebook, 'White Hat Programme'.

21 TikTok, 'TikTok Transparency'. [Accessed 5 March 2022].

22 Twitter, 'Twitter Transparency'. [Accessed 5 March 2022].

23 Facebook, 'Facebook Transparency'. [Accessed 5 March 2022].

24 Google, 'Google Transparency'. [Accessed 5 March 2022].

25 VKontakte, 'VK Safety'. [Accessed 5 March 2022].

26 Nytimes, 'Looking for Something New in Russia's 'New People' Party'. [Accessed 5 March 2022].

27 European Commission, 'Assessment of the Code of Practice on Disinformation', [Accessed 5 March 2022].

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.
Operating since 2014, we have carried out significant research enhancing NATO nations' situational awareness of the information environment and have contributed to exercises and trainings with subject matter expertise.

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