FALLING BEHIND:

HOW SOCIAL MEDIA COMPANIES ARE FAILING TO COMBAT INAUTHENTIC BEHAVIOUR ONLINE

PREPARED BY THE
NATO STRATEGIC COMMUNICATIONS
CENTRE OF EXCELLENCE

EXECUTIVE SUMMARY

From the 2014 invasion of Ukraine to more recent attempts to interfere in democratic elections, antagonists seeking to influence their adversaries have turned to social media manipulation.

At the heart of this practice is a flourishing market dominated by Manipulation Service Providers (MSPs) based in Russia. Buyers range from individuals to companies to state-level actors. Typically, these service providers sell social media engagement in the form of comments, clicks, likes, and shares.

Since its foundation, the NATO Strategic Communication Centre of Excellence in Riga has studied social media manipulation as an important and integral part of the influence campaigns malicious state and non-state actors direct against the Alliance and its partners.

To test the ability of Social Media Companies to identify and remove manipulation, we bought engagement on 105 different posts on Facebook, Instagram, Twitter, and YouTube using 11 Russian and 5 European (1 Polish, 2 German, 1 French, 1 Italian) social media manipulation service providers.

At a cost of just 300 EUR, we bought 3 530 comments, 25 750 likes, 20 000 views, and 5 100 followers. By studying the accounts that delivered the purchased manipulation, we were able to identify 18 739 accounts used to manipulate social media platforms.

In a test of the platforms' ability to independently detect misuse, we found that four weeks after purchase, 4 in 5 of the bought inauthentic engagements were still online. We further tested the platforms ability to respond to user feedback by reporting a sample of the fake accounts. Three weeks after reporting more than 95% of the reported accounts were still active online.

Most of the inauthentic accounts we monitored remained active throughout the experiment. This means that malicious activity conducted by other actors using the same services and the same accounts also went unnoticed.

While we did identify political manipulation—as many as four out of five accounts used for manipulation on Facebook had been used to engage with political content to some extent—we assess that more than 90% of purchased engagements on social media are used for commercial purposes.

We identified fake engagement purchased for 721 political pages and 52 official government pages, including the official accounts of two presidents, the official page of a European political party, and a number of junior and local politicians in Europe and the United States. The vast majority of the political manipulation, however, was aimed at non-western pages.

We further assessed the performance of the four social media companies according to seven criteria designed to measure their ability to counter the malicious use of their services. Overall, our results show that the social media companies are experiencing significant challenges in countering the malicious use of their platforms. While they are better at blocking inauthentic account creation and removing inauthentic followers, they are not doing nearly as well at combating inauthentic comments and views.

Based on this experiment and several other studies we have conducted over the last two years, we assess that Facebook, Instagram, Twitter, and YouTube are still failing to adequately counter inauthentic behaviour on their platforms.

Self-regulation is not working. The manipulation industry is growing year by year. We see no sign that it is becoming substantially more expensive or more difficult to conduct widespread social media manipulation.

In contrast with the reports presented by the social media companies themselves, our report presents a different perspective: We were easily able to buy more than 54 000 inauthentic social media interactions with little or no resistance.

Although the fight against online disinformation and coordinated inauthentic behaviour is far from over, an important finding of our experiment is that the different platforms aren't equally bad—in fact, some are significantly better at identifying and removing manipulative accounts and activities than others. Investment, resources, and determination make a difference.

Based on our experiment, we recommend:

1. **Setting new standards and requiring reporting based on more meaningful criteria**

2. **Establishing independent and well-resourced oversight of the social media platforms**

3. **Increasing the transparency of the social media platforms**

4. **Regulating the market for social media manipulation**

" Social media manipulation is the new frontier for antagonists seeking to influence elections, polarise public opinion, and side-track legitimate political discussions.

INTRODUCTION

Social media manipulation is the new frontier for antagonists seeking to influence elections, polarise public opinion, and side-track legitimate political discussions.

A new industry has developed to feed the market for inauthentic comments, clicks, likes, and followers. The first Manipulation Service Providers (MSPs) to cater to this new need appeared in Russia, but opportunistic MSPs soon began appearing in Europe, often simply reselling Russian-based services.

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 on-line to influence public perception of trends and popularity. Some inauthentic accounts are simple, bot-controlled [short for robot] accounts without pictures or content that view videos or retweet content following a computer program. Others are elaborate or '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. Bot-controlled accounts cost only a few cents each and are expected to be blocked quickly. More elaborate inauthentic accounts require some direct human control. They can cost up to several hundred euros and often remain online for years.

The 'Black market' for social media manipulation

The scale is greater than thought.
The infrastructure for developing and maintaining social media manipulation software, generating fictitious accounts, and providing mobile proxies is vast .

The openness of this industry is striking.
Rather than a shadowy underworld, it is an easily accessible marketplace that most web users can reach with little effort through any search engine. In fact, manipulation service providers advertise openly on major platforms.

Russian service providers dominate the social media manipulation market. Virtually all of the major manipulation software and infrastructure providers identified by us are of Russian origin.

The size of the social media manipulation industry is troubling. We have identified hundreds of providers. Several have many employees and significant revenue. It is clear that the problem of inauthentic activity is extensive.
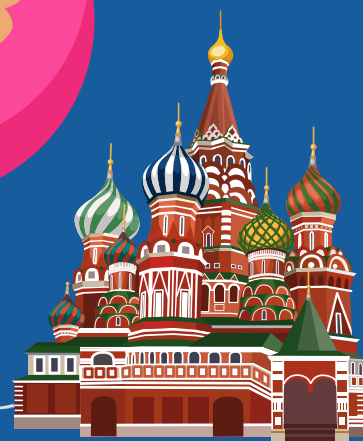
Commitments to prevent platform abuse

Social media companies have made several formal statements expressing their intent to tackle abuse of their platforms. The clearest formal commitment occurred in September 2018, when representatives of the major online platforms agreed on a self-regulatory Code of Practice to address the spread of online disinformation.[1]

One important part of the Code of Practice was a commitment to put into place clear policies for identifying and handling the misuse of automated bots, and to enforce these policies within the European Union.

The European Commission has urged social media companies to step up their efforts and, in view of the 2019 European elections, the Commission and the European Regulators Group for Audiovisual Media Services (ERGA) assessed the actions taken by Facebook, Google, and Twitter based on reports submitted by these platforms to the Commission.[2]

The Commission notes that the social media platforms reported that they had taken action against inauthentic behaviour to limit the scope of spam and disinformation globally.

Google reported to have globally removed more than 3.39 million Youtube channels and 8,600 channels for violations against its spam and impersonation policies. Facebook disabled 2.19 billion inauthentic accounts in the first quarter of 2019 and acted specifically against 1,574 non-EU-based and 168 EU-based pages, groups and accounts engaged in inauthentic behaviour targeting EU Member States. Twitter challenged almost 77 million spamlike or inauthentic accounts globally.[3]

These numbers are impressive, but themselves, they do no prove effectiveness. It is important to evaluate if the social media companies are truly living up to their commitments, and to independently verify their ability to counter misuse of their platforms.

In this report, we use experimental methods to identify how hard it is to circumvent the measures that should now have been adopted. Without access to data from the social media companies, we had to develop a much more creative approach to the problem.

Building on our previous work on the 'black market' for social media manipulation, we decided to use services selling inauthentic social media interactions to our benefit. The scale and effectiveness of the market for manipulation enables experiments to test and assess the ability of individual social media companies to counter manipulation.

Who we are

The NATO Strategic Communications Centre of Excellence is a multi-nationally 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 by multinational and cross-sector participants from the civilian, military, private, and academic sectors and from the use of modern technologies, virtual tools for analysis, research, and decision making.

NATO and the European Union (EU) are essential partners who have developed a closer cooperation to improve security for European citizens during the last few years—with joint declarations made to that effect in 2016 and 2018.[4] In Warsaw in July 2016, the two organisations outlined areas for strengthened cooperation in light of common challenges to the east and south. Areas of cooperation include countering hybrid threats, enhancing resilience, building defence capacity, maintaining and improving cyber defence.[5] Measures to bolster resilience to hybrid threats—from disinformation campaigns to emergent crises—are an essential part of NATO-EU cooperation today.

We developed this experiment in support of the European Union Action Plan against Disinformation[6] and the self-regulatory Code of Practice[7] to address the spread of online disinformation.

The malicious use of social media has shown to be an important tool for actors conducting influence activities against the interests of the EU and NATO. Bolstering our collective resilience requires a deeper understanding of this problem space so that we can establish effective analysis, prevention, and early detection. This will only be possible if we identify and address the vulnerabilities of social media platforms.

" Spending 300 EUR, we bought 3 530 comments, 25 750 likes, 20 000 views, and 5 100 followers, enabling us to identify 18 739 inauthentic accounts being used for social media manipulation.

SOCIAL MEDIA MANIPULATION EXPERIMENT

Introduction to the experiment

To test the ability of Facebook, Instagram, Twitter, and Youtube to identify and remove manipulation on their platforms we bought comments, views, and likes from a range of European and Russian social media manipulation service providers.

We structured the experiment so that we could measure and evaluate the performance of both manipulation service providers and social media platforms.

To limit the risk of unintentionally influencing real conversations online, we conducted the vast majority of the experiment by buying engagement on inauthentic profiles we created ourselves.

To assess if there is a difference between the various platforms' ability to counter bought manipulation on verified accounts, we also purchased comments and likes on a few real verified posts on each platform.

To make sure that we did not influence real conversations we only bought engagement on posts that were at least six months old and contained neutral apolitical messages such as New Year's greetings.

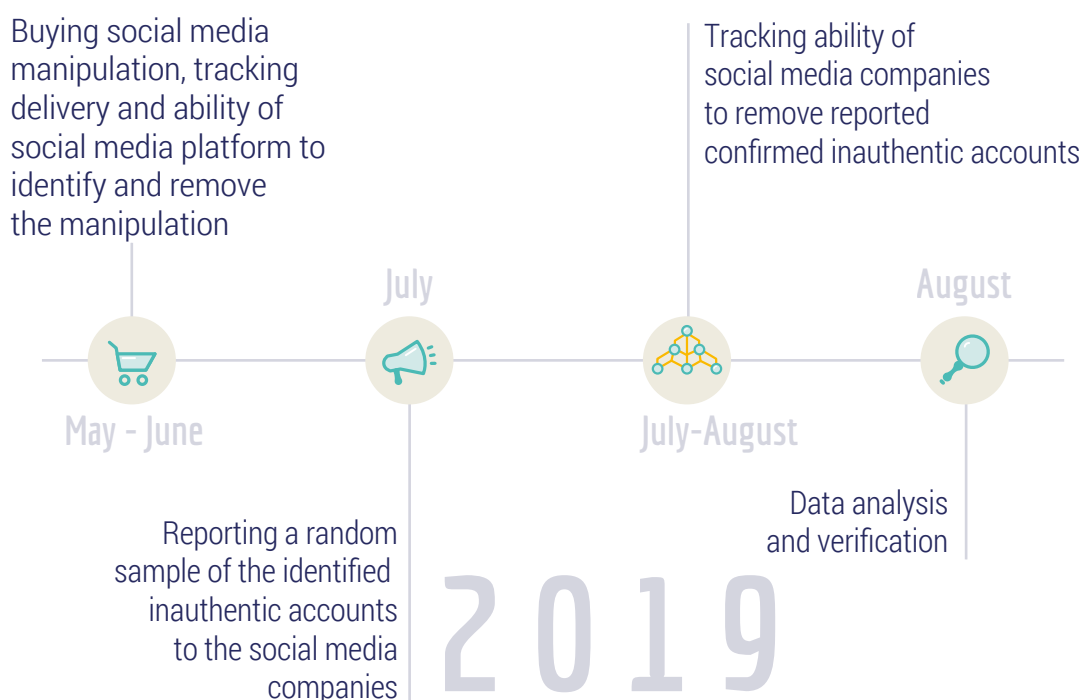The comments we bought were simple messages of a positive nature such as

'Hello!' and 'Thank you!' (see case-study on page 25). Engaging with posts that likely would not receive genuine engagement also enabled more accurate measurement of the purchased engagement.

As the context is also important, we chose to collect our data in the context of an election—a time when platforms had committed to be especially vigilant. The findings presented below, therefor, represent something of a best-case scenario for the social media companies as they had committed to dedicating extra resources to prevent abuse during this time period.

It could be argued that bought manipulation is more likely to be detected if it is placed on current content, but because we wanted to test the ability of the social media companies

to identify and block bought manipulation it was important that our experiment did not prompt users or account managers to report our activity to the social media companies as this would have "poisoned" our data. **We did not want to test the ability of social media managers or the public to detect and report inauthentic activity.**

We bought our engagement from commercial manipulation service providers. This means that failure to remove the inauthentic accounts and inauthentic engagement we bought means that malicious activity conducted by other actors using the same services and the same accounts also did not get removed. **Our experiment, therefore, offers insight into the ability of the social media companies to deal with the commercial manipulation industry.**

Buying social media manipulation, tracking delivery and ability of social media platform to identify and remove the manipulation

July

August

Tracking ability of social media companies to remove reported confirmed inauthentic accounts

May - June

July-August

Reporting a random sample of the identified inauthentic accounts to the social media companies

Data analysis and verification

2019

The scale of the experiment

To conduct the experiment we bought engagement on 105 different posts on Facebook, Instagram, Twitter, and YouTube using 11 Russian and 5 European (1 Polish, 2 German, 1 French, 1 Italian) social media manipulation service providers. Spending 300 EUR, we bought 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.

The experiment was carried out during six weeks in May and June 2019. To assess the ability of the platforms to remove the inauthentic engagement, we monitored the bought engagement from before engagement to one month after engagement. We reported the inauthentic engagement to the social media companies in July and continued monitoring through the end of August 2019 to measure the time it took for the social media platforms to react.

During the experiment, we recorded statistics on how quickly the manipulation service providers were able to deliver their services, and whether the quantity delivered was accurate. We then collected data on how the four social media platforms responded to the manipulated content by periodically measuring whether it had been removed.

The experiments were divided into several blocks of work, visualised below. In the following chapters we provide a detailed analysis of the ability of individual social media companies to detect and counter manipulation of their services.

Five steps of the experiment

01 Buying likes, comments, views, and followers for neutral posts on our own inauthentic accounts

02 Buying likes, comments, views, and followers for neutral apolitical posts

03 Tracking performance and response time of platforms in removing inauthentic activity

04 Tracking how long inauthentic accounts stay on the social media platform and with what they engage

05 Tracking how long it takes to remove accounts after reporting a random sample
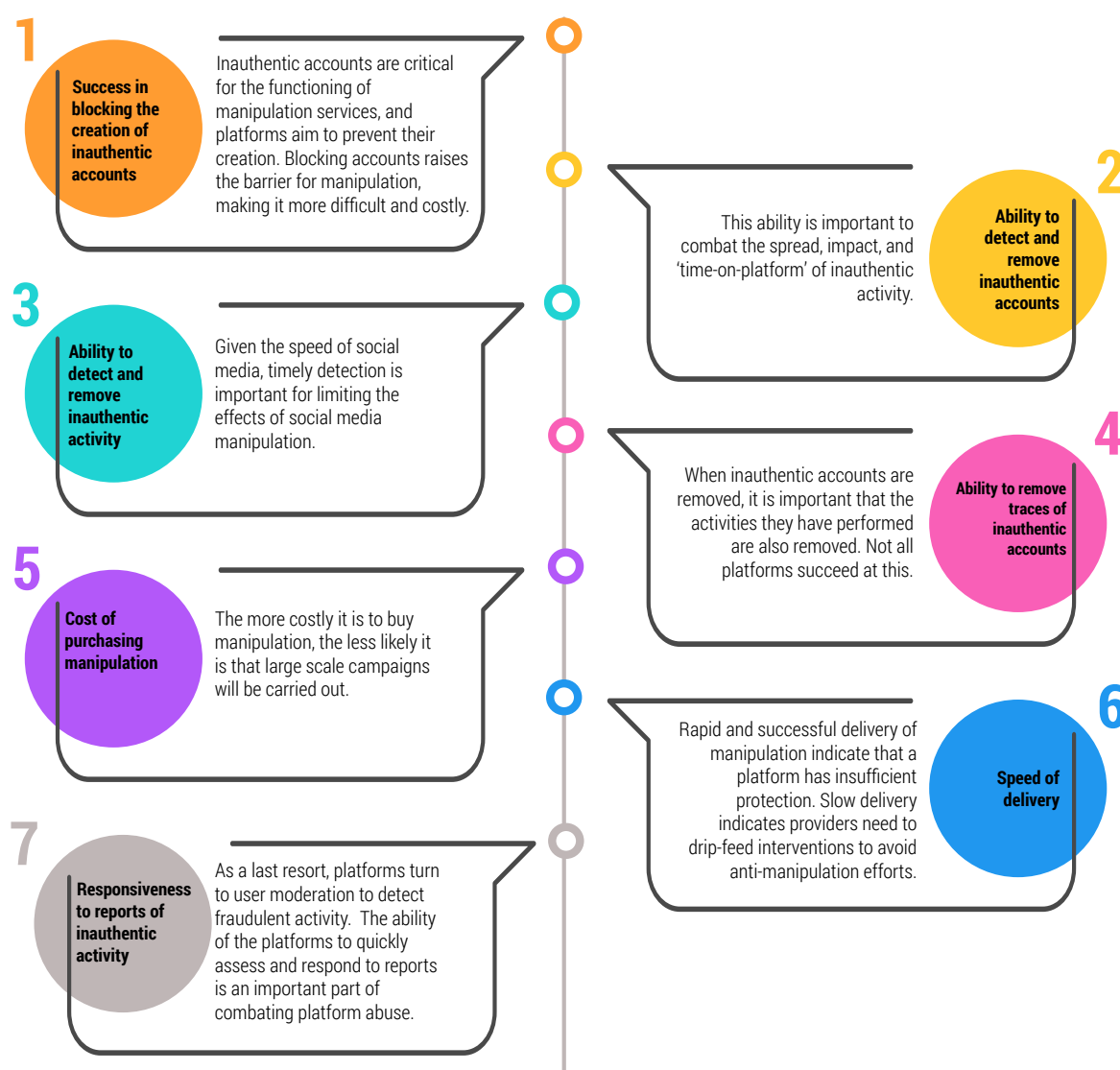
5 STEPS

Our assessment criteria

We assessed the performance of the four social media companies studied according to seven criteria designed to measure their ability to counter the malicious use of their services.

These criteria focus on detailed aspects of social media manipulation including blocking the creation of inauthentic accounts and the ability of platforms to recognise and remove coordinated inauthentic behaviour both independently and after such behaviour had been reported.

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

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 Ability to remove traces of inauthentic accounts

When inauthentic accounts are removed, it is important that the activities they have performed are also removed. Not all platforms succeed at this.

5 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.

6 Speed of delivery

Rapid and successful delivery of manipulation indicate that a platform has insufficient protection. Slow delivery indicates providers need to drip-feed interventions to avoid anti-manipulation efforts.

7 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.

Assesment of social media company ability to respond to inauthentic behaviour online

Blocking inauthentic account creation (Criteria 1)

In order to conduct the experiment, we had to set up our own inauthentic accounts. These accounts were used to upload content, which we then manipulated using Manipulation Service Providers (MSPs). Creating inauthentic accounts is becoming harder to do as some of the social media platforms have stepped up their efforts to combat inauthentic accounts. From our own pool of inauthentic accounts Facebook suspended 80 percent, Twitter—66 percent, and Instagram—50 percent. YouTube did not suspend any of our accounts.

By actively monitoring our own settings during account creation we were able to identify why we were blocked. The reasons varied from case to case, but included web browser cookies and the use of specific IP-addresses. Manually contacting support also allowed us to unblock accounts we needed for the experiment.

The systems used by Facebook, Instagram, and Twitter for protecting against the registration of multiple (inauthentic) accounts from a single IP address or a VPN are generally effective. The measures put into place by Twitter to prevent inauthentic account creation were especially hard to circumvent.

The experience of creating our own inauthentic accounts allows us to conclude that it is becoming far more difficult for the average user to create inauthentic accounts on the social media platforms we tested. However, the measures they use are not robust enough to stop persistent users or organisations. Of course, this is a step in the right direction, but more needs to be done to block inauthentic accounts from being created. YouTube especially needs to improve its efforts as it is by far the easiest platform to create inauthentic accounts on.

Countering manipulation (Criteria 2 - 4)

We assessed three criteria for the ability of social media companies to undo the effects of manipulation. First, the removal of bought activity, such as comments or likes. Second, the removal of the accounts used to deliver the manipulation. Finally, undoing all the effects created by inauthentic accounts. More advanced effect management could include notifications to account owners that their content has been manipulated, informing users who have been exposed to manipulation, and informing the public of significant cases.

While the different social media companies each have their strengths and weaknesses, one platform performs poorly no matter the criterion—Instagram was largely unable to detect and counter any manipulation. **Instagram managed to remove only 5% of bought inauthentic comments and virtually none of the inauthentic likes or views were corrected.**

YouTube is the only platform that succeeded in reducing manipulated view counts.

Across all platforms, the first decrease of purchased engagement is most often recorded on the third to the fifth day after the purchase, indicating that even when manipulation is removed it is often too slow to be effective.

Cost of purchasing manipulation (Criteria 5)

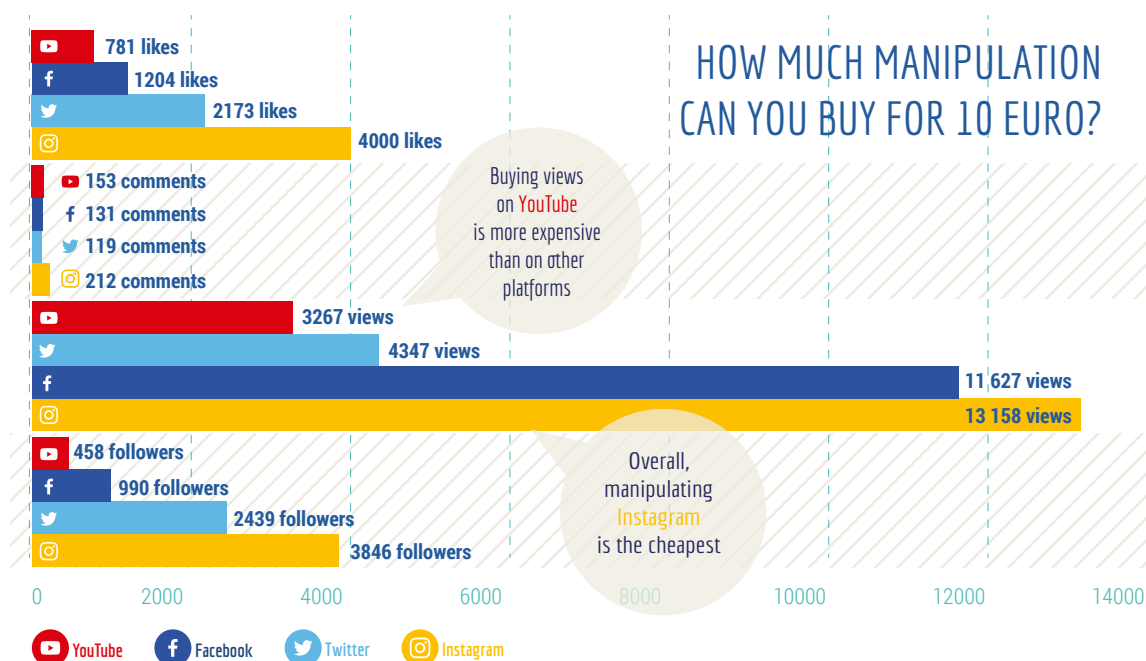To assess the cost of the services we sampled the offerings of five Russian providers.

The figure below show how many likes, comments, views, or followers it is possible to buy for 10 EURO. The firgures represent the average price quoted by five core Russian MSPs.

The cost of manipulation is a good indicator of how effectively social media platforms are at combating manipulation. When accounts used to perform manipulation are removed, MSPs have to spend time and money to replace them. When social media platforms redesign to break the scripts used to seed manipulation, MSP developers have to update their scripts. These costs are passed on to their consumers.

For only 300 EUR, we were able to buy 3 530 comments, 25 750 likes, 20 000 views, and 5 100 followers. A high proportion of this cost went to the more expensive and often less reliable European providers.

To make a balanced assessment of the relative cost of manipulating the platforms, we identified five inexpensive and reliable Russian MSPs that also provided a significant part of the manipulation services resold by other MSPs.

HOW MUCH MANIPULATION CAN YOU BUY FOR 10 EURO?

- 781 likes
- 1204 likes
- 2173 likes
- 4000 likes
- 153 comments
- 131 comments
- 119 comments
- 212 comments
- 3267 views
- 4347 views
- 11 627 views
- 13 158 views
- 458 followers
- 990 followers
- 2439 followers
- 3846 followers

Buying views on YouTube is more expensive than on other platforms

Overall, manipulating Instagram is the cheapest

YouTube   Facebook   Twitter   Instagram

Overall, YouTube is the most expensive service to manipulate. The cost of manipulation services for Twitter and Facebook is roughly similar, though it varies slightly depending on the service used. The cost of manipulation services for Instagram is nearly half that of the same manipulation for Twitter and Facebook and only a fifth of the cost of YouTube manipulation.

Manipulation of YouTube is the most expensive for everything except comments, whereas manipulation on Instagram is the cheapest in every category.

Speed of delivery
(Criteria 6)

Social media manipulation services are widely available, accessible, and professional. Almost all of the manipulation service providers we used were highly responsive to questions and complaints indicating that the manipulation service industry has managed to develop into a reasonably reliable industry.

We found that Instagram manipulation services overall were the most reliable, while comment services for Facebook, Twitter, and Youtube were the least reliable.

With the exception of Twitter, all the bought manipulation was delivered within 24 hours on average. While the average time of complete delivery on Instagram was also less than 24 hours, a number of service providers offered significantly faster services. Our record time from purchase to delivery was less

than five minutes on Instagram. Twitter was generally the slowest with bought manipulation first appearing after one hour and completing within two days on average.

There is a considerable amount of cross-activity in the manipulation industry. We noted that different European providers would often use the same inauthentic accounts, and these accounts were often of Russian and Ukrainian origin, indicating that many of the European MSPs use the same Russian sub-providers.

While most manipulation service providers are quite reliable, the volumes delivered were often not what had been bought. Sometimes we received fewer, but most often we received more. This may be because providers know some of their manipulation efforts will disappear—indeed at least one supplier offers a monthly guarantee, promising to periodically 'top-up' numbers if the social media company counters the manipulation.

However, we believe the real explanation is that many providers resell services offered by others and therefore are unsure exactly how many interventions will be delivered in a timely fashion.

This argument is supported by the considerable overlap in account use between providers. On Twitter, for instance, we found many examples of the same inauthentic accounts being used by five or more separate MSPs.
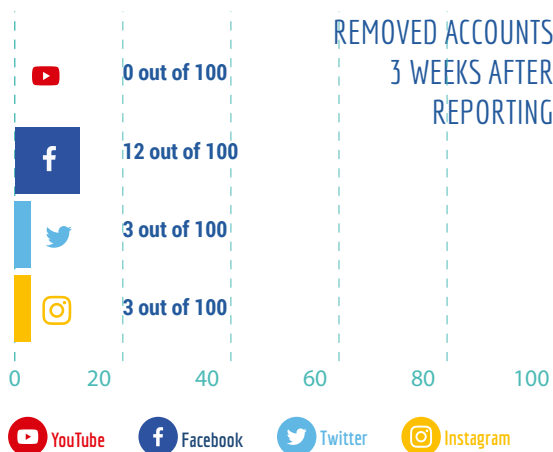
Responsiveness (Criteria 7)

After the end of phase one of the experiment, we reported 100 random accounts used for social media manipulation for each platform and then we monitored how long it took for the platforms to remove the accounts. It is worth reiterating that the accounts we reported were the accounts that delivered the manipulation we bought, meaning that we were 100% certain that these accounts were engaging in social media manipulation.

Three weeks from the date we reported the accounts, the social media platforms had only removed 4,5% of the accounts we had reported to them.

Given the low number of accounts removed it is clear that social media companies are still struggling to remove accounts used for social media manipulation, even when the accounts are reported to them.

REMOVED ACCOUNTS 3 WEEKS AFTER REPORTING

YouTube — 0 out of 100
Facebook — 12 out of 100
Twitter — 3 out of 100
Instagram — 3 out of 100

YouTube   Facebook   Twitter   Instagram

The social media manipulation industry

This experiment has strenghtened many of the conclusions from our report on the 'black market' for social media manipulation.[8]

The manipulation market is widely available, and there is a large degree of reselling, which means that different providers often use the same set of accounts for their manipulation services. Some of the best MSPs are transparent with the size of the underlying set of inauthentic accounts and the quality of the service they provide. The worst providers simply pocket the money received without delivering any services.

But even if the market is somewhat chaotic, it functions reasonably well and most orders are delivered in a timely and accurate manner. Social media manipulation remains widely available, cheap, and efficient.

During platforms updates, the manipulation services usually stop functioning for a few days, but so far, they have always been able to circumvent new safeguards by the platforms and resume service within a week or so. It is clear that so far the social media platforms have been mostly unable to prevent the MSPs from abusing their platforms.

" While YouTube is the worst at removing inauthentic accounts, it is best at countering inauthentic likes and artificial video views.

ASSESSMENT OF THE PLATFORMS RELATIVE STRENGTHS AND WEAKNESSES

Assesment of platforms

Facebook

Facebook was the platform that was most successful at blocking inauthentic account creation. Facebook has sophisticated anti-automation systems built into the structure of the platform, and several MSPs struggled to offer consistent services. In some cases, otherwise reliable vendors were unable to deliver the promised manipulation on Facebook.

However, vendors who were able to circumvent Facebook's counter-measures had a very high success rates. Even after many weeks, it was rare for any of the inauthentic interactions to have disappeared. While they did better than Instagram by removing roughly 10 % ter than Instagram by removing roughly 10 %

of the inauthentic likes after a month, they did worse than Instagram by removing 0% of the inauthentic comments. It is also noteworthy that both Facebook and Instagram were especially weak at countering inauthentic video views. No inauthentic views were removed by the platforms, and the cost of inauthentic views on the platforms is disproportionately low compared to Twitter and YouTube.

Thus Facebook resembles a fortress with formidable defences facing the outside world, but qualified actors are still able to scale the walls of Facebook; policing and oversight within the walls is far less effective.

Instagram

Instagram was somewhat successful at blocking account creation with roughly a 50%

block rate, however it is quite easy to over-come their blocking by using relatively simple techniques such as VPNs and cache control.

The cost of manipulating Instagram was the lowest for all types of manipulation—likes, views, comments, and followers. Manipulation service providers found Instagram to be the easiest platform to manipulate. On average, the service provided deviated by 18% from what we ordered and most orders were delivered within 24 hours.

Instagram seemed to have a flaw in their system as their counters for likes and comments did not reflect real changes—during our experiment the counters went up but never down. When inauthentic accounts were removed there was no change in the like and comments counters. To get accurate recordings we had to compare the list of user engagements and the counter numbers for each engagement. A consequence of this flaw is that manipulating Instagram is easier because even if your bots are blocked the effect of the manipulation will remain on the platform.

Instagram also performed poorly in blocking content manipulation. Instagram removed only one percent of the bought likes during the month-long test phase. Instagram also had the lowest number of blocked accounts—14%—by the time we started reporting the inauthentic account to the social media platforms. Instagram did better than Twitter by blocking 44 % of the inauthentic followers we bought, but worse than YouTube.

During our experiment phase Instagram launched a significant upgrade of their platform, which caused some of the manipulation service providers to pause their services, but within a few weeks all the MSPs had updated their system and were able to resume their manipulation services.

Our experiment clearly shows that Instagram has significant challenges with countering abuse on their platform as manipulating Instagram it is both easy and cheap.

Twitter

Twitter is currently the most effective platform at countering abuse of their services. It takes longer for bought engagement to appear on Twitter and the quality of delivery is more uneven than on the other platforms. Even so, all the MSPs delivered all the services we bought without any refusals or failed deliveries.

Twitter also identified and removed more manipulation than the other platforms. On average half of the likes and retweets bought on Twitter were removed during the testing period. At 35%, Twitter had blocked the highest proportion of accounts by the time we started reporting the accounts. This indicates that accounts used by MSPs are removed most effectively on Twitter.

Twitter failed to remove any of the bought views and we were unable to measure the number of comments removed because of a Twitter feature that made it difficult for us to measure the reason a comment was removed

[the much debated "This tweet is unavailable" feature].

While Twitter is effective at blocking new inauthentic accounts, the legacy of their inadequate anti-spam efforts still impacts the platform today as aged inauthentic accounts that were created before Twitter improved their defences seem to remain active on the platform.

Twitter should be commended for doing the most to combat malicious use of their platform. They are currently ahead of the other platforms we tested. Even so, it is still very possible to manipulate content and conversations on Twitter, it just requires a little more effort than on the other platforms.

YouTube

Our assessment of YouTube shows a split picture. **While YouTube is the worst at removing inauthentic accounts, it is best at countering inauthentic likes and artificial video views.** YouTube's ability to counter inauthentic comments is twofold; while many of the manipulation providers struggled to provide service, one provider was extremely efficient in delivering inauthentic comments on YouTube. Nine out of ten comments delivered remained active on the platform throughout the experiment.

Countering artificial views should be the greatest concern for YouTube as fake views generate fake advertising costs for YouTube advertisers. **Based on our experiment YouTube is the industry leader in countering artificial**

views, however a 10 % reduction is far from sufficient for preventing platform abuse. From previous experiments we have seen that inauthentic activity on YouTube can remain active for many months without being detected, an insight this experiment seems to strengthen.

In many ways, YouTube is the least transparent platform, and it is difficult to identify inauthentic accounts on YouTube. The popularity of the platform, the difficulty [for external researchers] of detecting platform manipulation, and the potential financial rewards of manipulation make YouTube an ideal target. The fact that YouTube is the most expensive platform to manipulate is either a testament to its defensive actions or to the popularity of YouTube manipulation. We currently do not know which.

Relative performance

There is a significant difference between the ability of the different social media platforms to counter manipulation.

While Twitter outperformed the others, it is far more challenging to rank the relative performance of YouTube, Facebook, and Instagram. In our final assessment we decided to prioritise a platform's ability to counter manipulation before other responses, which places YouTube ahead of Facebook and Instagram. At the same time, we assess, that none of the four platforms are doing enough to prevent the manipulation of their services.

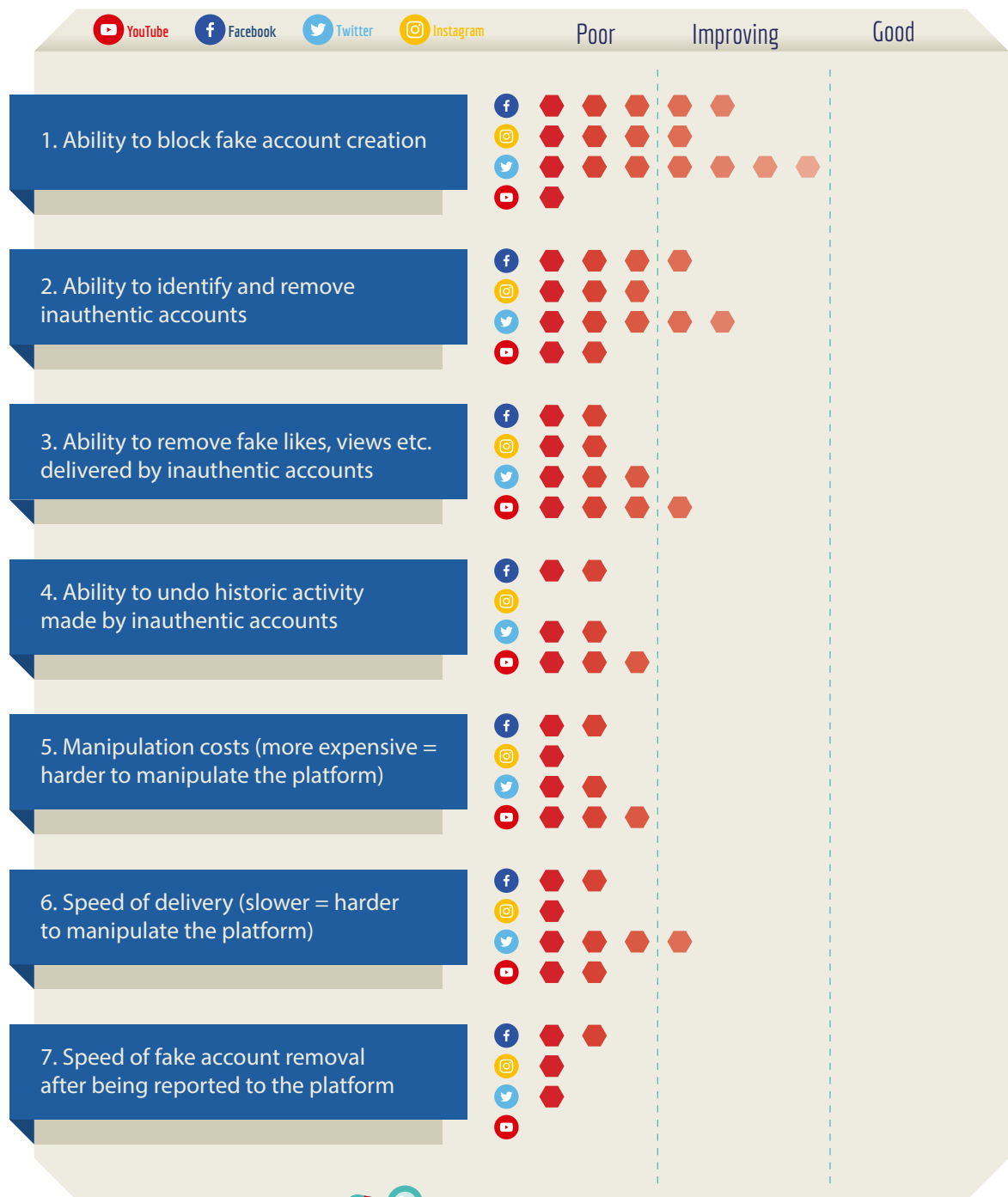# ASSESSMENT OF THE PLATFORMS RELATIVE STRENGTHS AND WEAKNESSES

YouTube   Facebook   Twitter   Instagram

Poor | **Improving** | **Good**

1. Ability to block fake account creation

2. Ability to identify and remove inauthentic accounts

3. Ability to remove fake likes, views etc. delivered by inauthentic accounts

4. Ability to undo historic activity made by inauthentic accounts

5. Manipulation costs (more expensive = harder to manipulate the platform)

6. Speed of delivery (slower = harder to manipulate the platform)

7. Speed of fake account removal after being reported to the platform

Illustration of relative performance of Twitter (1st), YouTube (2nd), Facebook (3rd) and Instagram (4th). Manipulation service providers are still winning.

The industry's ability to remove manipulation

Overall social media companies are experiencing significant challenges in countering the malicious use of their platforms. While they are improving in the field of removing inauthentic followers, they are facing substantial difficulties in combating inauthentic comments and views.

Across all platforms, removal of purchased manipulations is most often first recorded on the third to the fifth day after purchase, which is worrying given the speed of social media. If manipulations are identified and removed only three to four days after they are posted on Instagram, Facebook and Twitter, the delayed efforts to counter manipulation will be less effective.

If we focus on the average reduction of manipulation per post after four weeks, then the relative performance of the different platforms is the same as the total decrease seen in the table below. Twitter - 60%, YouTube - 43%, Facebook - 26%, and Instagram - 1%.

Industry's ability to remove inauthentic accounts

Six weeks after we started buying inauthentic social media engagement, just before we started reporting the accounts used to deliver the manipulation services, we measured how many were still active. The results are disturbing.

In total, just 17% of the bots we identified had been removed. This low figure shows that the social media companies' own algorithms for detecting misuse are ineffective across the board. The worst-performing services for blocking inauthentic accounts were Instagram and YouTube. Facebook ranked third, having removed 21%. And, according to this measure, the least-poorly-performing service was Twitter, which succeeded in removing 35% of the accounts.

These figures bear consideration: **bad actors wishing to manipulate social media plat-**

	Comments	Likes	Followers	Views
Instagram	3%	1%	44%	0%
Facebook	0%	10%	n/a	0%
Twitter	1%	27%	37%	0%
YouTube	11%	30%	61%	10%

Percentage of inauthentic engagement removed after four weeks.

forms can expect that only a small fraction of inauthentic activity will be proactively removed. So, most malicious activity goes unchallenged, and even if it is removed, its effects often remain.

Social media companies report that they block millions of inauthentic accounts annually,[9] but this does not seem to influence the ability of manipulation service providers to manipulate their platforms. One explanation for the seemingly impressive numbers reported by the social media companies could be that millions of accounts are blocked upon creation.

If malicious actors fail repeatedly before successfully creating inauthentic accounts, the result is that a high number of inauthentic accounts are blocked but inauthentic accounts are still eventually created.

Inauthentic accounts pattern of interaction on social media

Following the accounts used for manipulation to analyse who else uses their services and which content they are manipulating is technically difficult due to data access limitations.

While we were able to examine accounts on Twitter, on Facebook we were only able to see what pages the accounts had engaged with and on Instagram we had to conduct a manual examination of a random sample of accounts. We found no way to study this issue on YouTube.

Our experiment shows that the vast majority of bought engagement is used for commercial purposes. Instagram seems to have the biggest problem with bought manipulation on

14% Instagram

21% Facebook

35% Twitter

11% YouTube

0 1500 3000 4500 6000 7500 9000 10500 12000 13500 15000

Instagram: 14 330 accounts identified
Facebook: 2427 accounts identified
Twitter: 1982 accounts identified
Youtube: 471 accounts identified

● Blocked

○ Initial number
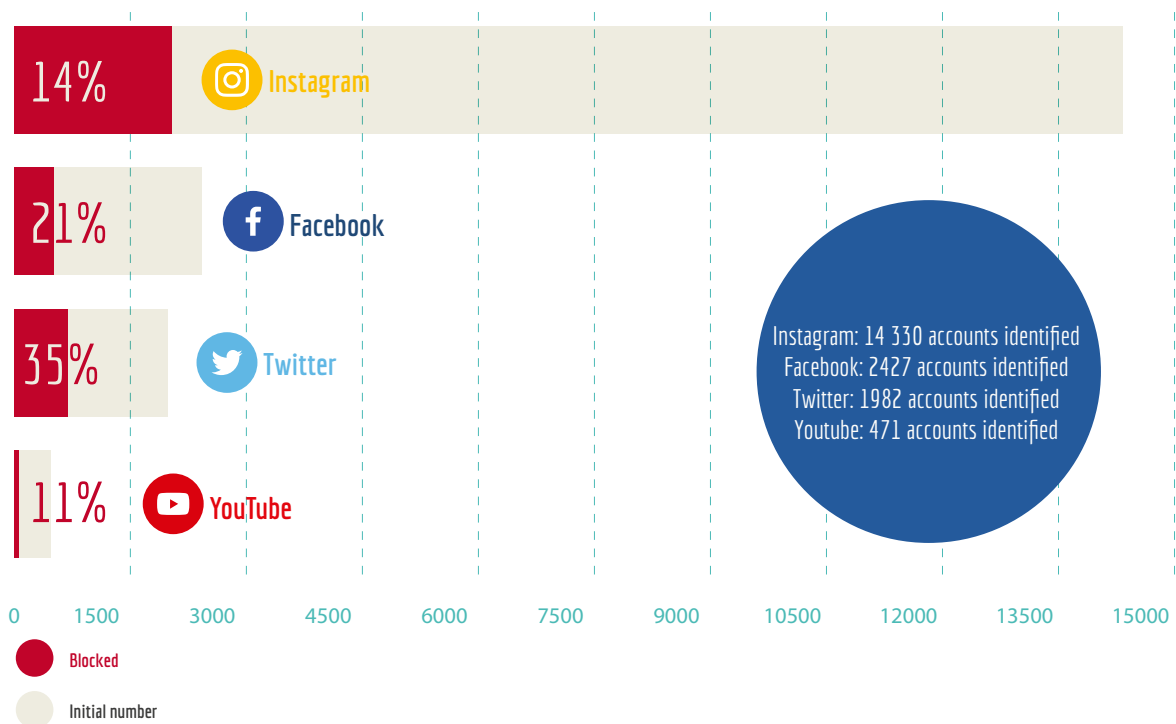
Percentage of inauthentic accounts removed after six weeks.[10, 11]

commercial influencer accounts. Some of the influencers we identified even had contracts with major international brands and were manipulating their reach and engagement statistics.

While we did identify political manipulation, and as many as four of five accounts used for manipulation on Facebook had been used to engage with political content to some extent, we assess that more than 90% of purchased engagements on social media are used for commercial purposes.

At the same time, it should be noted that we did identify at least one known pro-Kremlin bot account in our pool of identified inauthentic accounts. **This indicates that even if political manipulation is only a minor function of the manipulation industry, it is definitely being used for this purpose as well.** The inauthentic accounts we identified had been used to buy engagement on 721 political pages and 52 government pages, including the official accounts of two presidents, the official page of a European political party, and a number of junior and local politicians in Europe and the United States. The vast majority of the political manipulation, however, was aimed at non-western pages.

CASE STUDY: PROTECTING VERIFIED ACCOUNTS

While most of our experiment was conducted by buying engagement on our own accounts we also tested if the platforms are better at protecting verified institutional accounts.[10]

To conduct this part of the experiment we bought engagement on apolitical messages such as New Year's greetings on the posts of a few European institutional accounts. The comments we bought were simple messages of a positive nature such as 'Hello!' and 'Thank you!'

We have chosen four posts by commissioners Jourová, Katainen and Vestager to illustrate this case study. The results are representative for our overall conclusion that engagement is quickly delivered and remains active for a significant time period. Our examples show that inauthentic activity can remain active for a long time. In December 2019, 30 weeks after the experiment, a significant proportion of the inauthentic content was still online.
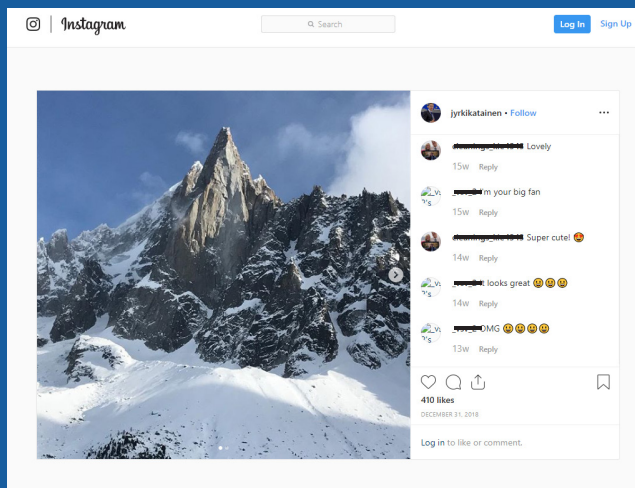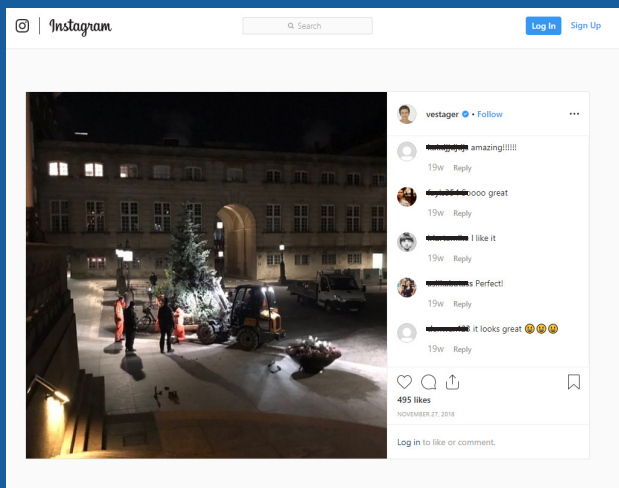
These four examples also show that sometimes manipulation service providers over deliver, sometimes they fail, but mostly they are right on target. Most often it is the quality of the manipulation service provider, not the platform, which determine effectiveness.

We had predicted that verified accounts and posts would be better protected, but based on our case-study, this does not seem to be the case. We therefore conclude that verified accounts, even institutional accounts, are no better protected against manipulation than any other accounts on social media platforms. More specifically, we were unable to detect any difference between the protection of verified and ordinary accounts on Instagram, but for Facebook there was a difference between the reductions in inauthentic likes. Verified accounts had an average of 27% of their likes removed after four weeks while our ordinary accounts only lost 5%. For inauthentic comments and views there was no difference on Facebook.

On Twitter, there was a clear difference between bought engagement on our ordinary posts versus bought engagement on verified posts, but without any clear pattern. 37% of comments on our own posts were removed, but only 1% of inauthentic comments were removed from the verified posts. For likes and retweets Twitter was more effective at reducing inauthentic engagement on verified posts by roughly 20%. On YouTube we saw a 23% removal of inauthentic likes on verified videos versus a 1% decrease on ordinary videos. However, for comments the results were reversed with an 13% removal rate on ordinary videos compared to a 8% decrease on verified videos.

So while we did identify a difference in the ability of social media platforms to protect verified European institutional accounts, the identified difference was either random or insignificant. It leaves us to conclude that verified institutional accounts are likely no better protected against manipulation on social media platforms.
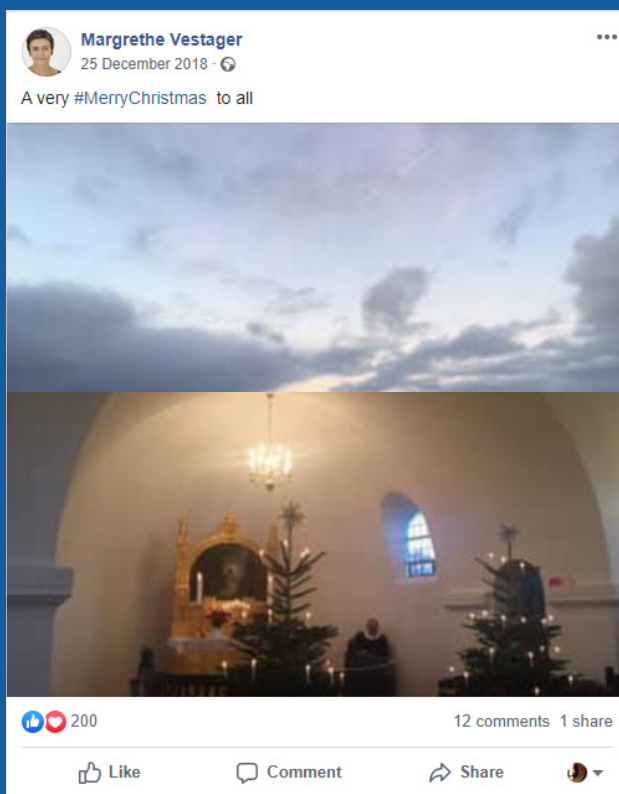
Margrethe Vestager
25 December 2018 · 🌐

A very #MerryChristmas to all

❤ 200

12 comments 1 share

👍 Like   💬 Comment   ↪ Share

Věra Jourová ✓
@VeraJourova

Merry Christmas to all my followers! 🎅🎄👼

from
EU Justice
and Consumers
Merry Christmas

European Commission

10:53 AM · Dec 24, 2018 · Twitter for iPhone

55 Retweets   **76** Likes

" Self-regulation is not working. The manipulation industry is growing year by year. We see no sign that it is becoming substantially more expensive or more difficult to conduct widespread social media manipulation.

CONCLUSIONS

Since its foundation, the NATO Strategic Communication Centre of Excellence has studied social media manipulation because it is an important and integral part of the influence campaigns malicious state and non-state actors direct against the Alliance, Allied nations, and Partner nations. Bolstering resilience to influence campaigns is an essential part of what we do.

In `tackling online disinformation: a European approach`, the European Commission rightly noted that:

> [T]he exposure of citizens to large scale disinformation, including misleading or outright false information, is a major challenge for Europe. [...]

[N]ew technologies can be used, notably through social media, to disseminate disinformation on a scale and with speed and precision of targeting that is unprecedented, creating personalised information spheres and becoming powerful echo chambers for disinformation campaigns. [...] Mass online disinformation campaigns are being widely used by a range of domestic and foreign actors to sow distrust and create societal tensions, with serious potential consequences for our security.[12]

In this context, it is vitally important that social media companies do their utmost to prevent the abuse of their platforms. We de-

signed this experiment to test the leading social networks' implementation of the part of the self-regulatory Code of Practice which addresses inauthentic accounts and coordinated inauthentic behaviour online.

Based on this experiment and several other studies we have conducted over the last two years, we assess that Facebook, Instagram, Twitter, and YouTube are still failing to tackle coordinated inauthentic behaviour online.

Self-regulation is not working. The manipulation industry is growing year by year and we see no signs that conducting widespread social media manipulation is becoming substantially more expensive or more difficult.

We have followed the reports of the social media companies as delivered within the framework of the European self-regulatory Code of Practice. We recognize that all the platforms have undertaken efforts to address coordinated inauthentic behaviour, fake accounts, and malicious, bot-driven activity as well as terms of service enforcement during the past year. At the same time, it is evident that the level of transparent and independent assessments that would enable accurate conclusions are still missing.

In contrast to the reports presented by the social media companies themselves, we offer a different picture: We were easily able to buy more than 54 000 inauthentic social media interactions with little or no resistance. Our experiment shows that social media platforms can still be easily manipulated.

Most of the inauthentic accounts we monitored remained active throughout the experiment. This means that malicious activity conducted by other actors using the same services and the same accounts also went unnoticed.

This means that even if our experiment, which was benign by design, did not trigger our purchased engagement to get blocked – neither did any of the other activity performed by the leading social media manipulation companies in Europe and Russia. We know this because the bulk of the inauthentic accounts used to deliver the engagement we bought were also used to deliver engagement other customers bought – and the inauthentic accounts mostly stayed active during our entire test. In fact, many are active still today.

Although the fight against online disinformation and coordinated inauthentic behaviour is far from over, an important finding of our experiment is that the different platforms aren't equally bad—in fact, some are significantly better at identifying and removing manipulative accounts and activities than others.

Investment, resources, and determination make a difference.

❞ If we can conduct an experiment, then so can the social media companies.

POLICY RECOMMENDATIONS

Set standards and require reporting based on more meaningful criteria

To further evaluate the impact and extent of inauthentic activity on social media platforms, more granular information is needed on the kind of inauthentic accounts blocked, which kind manages to gain access to the platforms and what impact they are having. More detailed insights are also required about detected inauthentic coordinated activity, including targets, levels of engagement, and the issues exploited to manipulate public opinion. Furthermore, a common standard needs to be developed so that reports from different social media companies can be compared to

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

Increase transparency

If we can conduct an experiment, then so can the social media companies. We faced significant challenges because we were forced to collect snippets of information from the outside, but the companies could test their defences and report the results with much greater accuracy if they chose to do so. Currently, they are mostly reporting good news—such as how successful they have been in preventing the creation of inauthentic accounts. In doing so, they present a picture that provides

too little insight into how many inauthentic accounts eventually gain access and what they do on the platforms.

More transparency is needed to understand the scope and effect of manipulation. Meta-manipulation, the practice of buying engagement to trigger algorithms to boost posts, is especially worrying since it is very difficult for outside researchers to identify.

Establish independent and well-resourced oversight

Independent oversight could be able to provide the insight needed to better assess the progress of the social media companies in countering inauthentic activity on their platforms. Today we are in a situation where efforts to analyse, evaluate and assess social media companies are facing a resource disadvantage. Data is becoming scarcer and our opportunities to research this field is constantly shrinking. This effectively transfers the ability to understand what is happening on the platforms to social media companies. Independent and well-resourced oversight is needed.

Regulate the market for social media manipulation

While we have focused a great deal on the ability of the 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 regulate the market for social media manipulation.

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

Manipulative service providers continue to advertise and promote their services on the very platforms, which they seek to undermine. Providers trafficking in YouTube manipulation services buy ads from Google—the owner of YouTube—and fearlessly promote their services using both advertisements and YouTube channels. It is far too easy to find and order manipulation services on the very platforms they seek to undermine.

WhatsApp, a company owned by Facebook, issued a stern warning in June 2019 noting that "[…] beginning on December 7, 2019, WhatsApp will take legal action against those we determine are engaged in or assisting others in abuse that violates our Terms of Service.[…]".[13] As manipulation service providers engage in the systematic abuse of social media companies it is surprising that there has not been more systematic efforts by the social media companies to counter the manipulation industry. The announcement by WhatsApp is an important step forward.

A whole-of-industry solution is needed

Recent studies have shown that social media manipulation and disinformation generate significant ad-revenue.[14] In fact, a recent report argues that inauthentic influencer mar-

keting is a 1,3 billion dollar per year problem.[15] Our research confirms that commercial manipulation is indeed the main driving force for social media manipulation—gaming advertisers for profit. But the tools and methods developed and funded to scam the advertising industry are also used for political and national security interference. At the same time the telecommunication industry has a responsibility to limit the use of sim-cards for manipulation services, and most manipulation service providers depend upon financial payment solutions (such as Paypal) to function well, and they require Internet Service Providers to gain and maintain access to the internet. A whole-of-industry solution is needed to combat this problem.

Implications for NATO

Social media manipulation is a challenge for NATO because it is an important tool for malicious actors conducting influence activities against the interests of the Alliance. Bolstering our collective resilience requires us to understand this problem better so we can establish more effective procedures for analysis, prevention, and early detection. 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 malign purposes.

If antagonists are able to manipulate the information environment, the ability of the Alliance to effectively message in times of crisis or conflict will be hampered. Therefore, the Alliance must continue to refine its strategies for communication in a highly contested Information Environment.

Assessing the Information Environment requires a further refined ability to differentiate between real and genuine content. The findings of this study should be incorporated into the Alliance's continued efforts to enhance its ability to assess the Information Environment.

ENDNOTES

1. European Commission. "Code of Practice on Disinformation." September 26, 2018.
2. EEAS - European External Action Service - European Commission. "Progress Report on Action Plan against Disinformation"
3. Ibid.
4. EEAS - European External Action Service - European Commission. "EU-NATO Cooperation - Factsheets"; NATO. "Relations with the European Union"
5. NATO. "Joint declaration by the President of the European Council, the President of the European Commission, and the Secretary General of the North Atlantic Treaty Organization".
6. European Commission - European Commission. "Action Plan on Disinformation: Commission Contribution to the European Council (13-14 December 2018)".
7. European Commission. "Code of Practice on Disinformation" September 26, 2018.
8. NATO StratCom CoE. "The Black Market for Social Media Manipulation".
9. European Commission. "Annual Self-Assessment Reports of Signatories to the Code of Practice on Disinformation 2019".
10. For YouTube, we are reporting the percentage of comments removed as a proxy for users. This estimate errs on the generous side, as comments can be moderated without a user being permanently banned from the platform.
11. The composition of bought engagement, and the type of inauthentic accounts used by the manipulation service providers, varied for each social media company.
12. European Commission. "Communication - Tackling Online Disinformation: A European Approach".
13. WhatsApp.com. "WhatsApp FAQ - Unauthorized Use of Automated or Bulk Messaging on WhatsApp".
14. Global Disinformation Index, "The Quarter Billion Dollar Question: How is Disinformation Gaming Ad Tech?".
15. CHEQ. "Ad Fraud 2019: The Economic Cost of Bad Actors on the Internet".

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