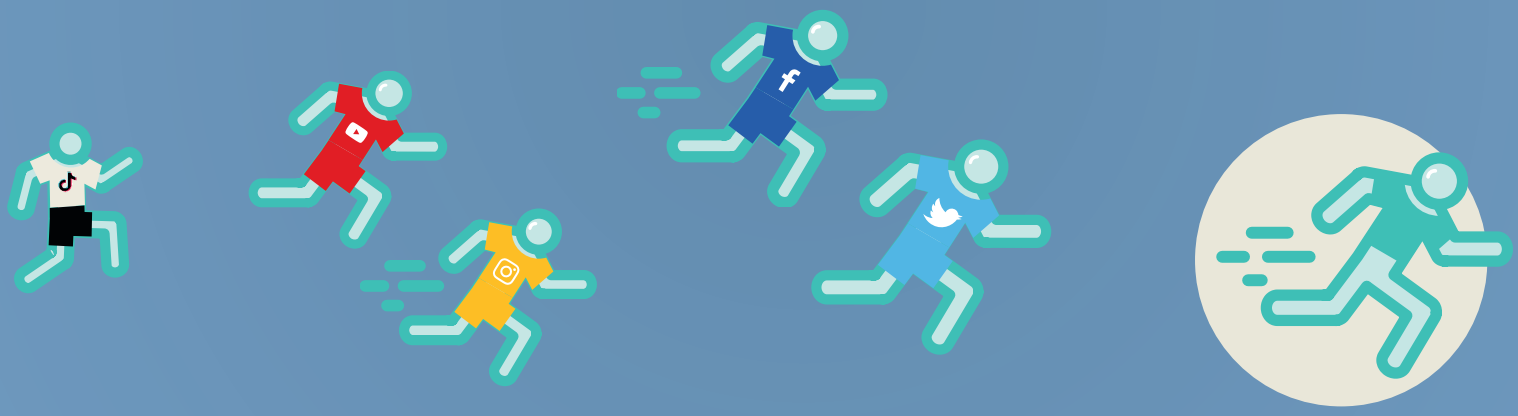
978-9934-564-92-5



SOCIAL MEDIA MANIPULATION 2020

HOW SOCIAL MEDIA COMPANIES ARE FAILING TO COMBAT INAUTHENTIC BEHAVIOUR ONLINE

PREPARED BY THE
NATO STRATEGIC COMMUNICATIONS
CENTRE OF EXCELLENCE

ISBN: 978-9934-564-92-5
Project Manager: Rolf Fredheim
Authors: Sebastian Bay, Anton Dek, Iryna Dek, Rolf Fredheim
Research: Singularex
Copy Editing: Anna Reynolds
Design: Kārlis Ulmanis

This report was completed in November 2020, based on an experiment that was conducted September to October 2020.

Singularex is a Social Media Intelligence and Analytics company based in Kharkiv, Ukraine.
Website: www.singularex.com
Email: hello@singularex.com

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

EXECUTIVE SUMMARY

Introduction

Antagonists, from foreign governments to terror groups, anti-democratic groups, and commercial companies, continually seek to manipulate public debate through the use of coordinated social media manipulation campaigns. These groups rely on fake accounts and inauthentic behaviour to undermine online conversations, causing online and offline harm to both society and individuals.

As a testament to the continued interest of antagonists and opportunists alike to manipulate social media, a string of social media companies, researchers, intelligence services, and interest groups have detailed attempts to manipulate social media conversations during the past year. Therefore, it continues to be essential to evaluate whether the social media companies are living up to their commitments to counter misuse of their platforms.

In an attempt to contribute to the evaluation of social media platforms, we re-ran our ground-breaking experiment to assess their ability to counter the malicious use of their services. This year we spent significant effort to improve our methodology further, and we also added a fifth social media platform, TikTok, to our experiment.

The experiment

To test the ability of social media companies to identify and remove manipulation, we bought engagement on thirty-nine Facebook, Instagram, Twitter, YouTube, and TikTok posts, using three high-quality Russian social media manipulation service providers. For 300 € we received inauthentic engagement 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.

While measuring the ability of social media platforms to block fake account creation, to identify and remove fake activity, and to respond to user reports of inauthentic accounts, we noted that some of the platforms studied had made important improvements. However, other platforms exhibited a continued inability to combat manipulation. Of the 337 768 fake engagements purchased, more than 98 per cent remained online and active after four weeks, and even discounting fake views, more than 80 per cent of the 14 566 other fake engagements delivered remained active after a month.

While these numbers aren't much to celebrate, we did observe significant pushback, primarily from Facebook and Twitter. In some cases, Facebook managed to remove as much as 90 per cent of the fake views before the manipulation service providers restored their work. This is an important step in the right direction.

Yet another win this year was a significant improvement by Facebook in blocking the creation of inauthentic accounts. But despite being the most effective at this, Facebook was the only platform studied to show an increasing half-life for active inauthentic accounts. While none of the measures introduced by any of the platforms are robust enough to stop persistent users or organisations from manipulation, both Facebook and Twitter have made it significantly more challenging.

Twitter stands out in its ability to remove inauthentic accounts from the platform; fake accounts disappear 40 per cent faster than in 2019, indicating that Twitter is three times faster than Facebook at removing accounts engaged in inauthentic activity.

Twitter and Facebook are now relatively good at blocking simpler automatic manipulation services, most notably automated comments, pushing manipulation service providers to rely on human labour to a greater extent. While there isn't any clear indication that this is driving up prices yet, any human-based manipulation will be more expensive than an automatic solution.

YouTube maintains its position as the least effective of the four major platforms we tested last year, and we have not seen any meaningful improvements since then. In fact, since our first report, YouTube has been overtaken by both Facebook and Instagram.

TikTok, the latest platform added to our experiment, seems to be nearly defenceless against platform manipulation. None of the manipulation was prevented or removed by the platform, making it by a distance the easiest to manipulate. In hindsight, we now realise that the other platforms could be doing worse.

Only YouTube can counteract fake views in any significant way. On Facebook, Instagram, Twitter, and TikTok, fake views continue to be quickly delivered, and are seemingly never removed. On Instagram, we managed to get 250 000 fake views delivered within an hour. Antagonists use fake views to manipulate platform algorithms to influence what social media users see. This remains a significant challenge for all platforms.

Our investigation has also shown that reporting an account for confirmed social media manipulation does not induce a platform to block that account. Of the 499 accounts (100 to Facebook, 100 to Instagram, 100 to YouTube, 100 to Twitter and 99 to TikTok) we reported as verified to have engaged in inauthentic behaviour, 482 accounts remained active five days after they were reported. Indeed, in all cases where we received feedback on the accounts we reported, it was to state that the user in question did not violate the platform's terms of service. We conclude that reporting and moderation mechanisms must be

improved so that a larger share of inauthentic accounts reported to the platforms are removed—even if only a single user reports them. Reported inauthentic accounts must not regularly escape sanction.

Conclusions

The most important insight from this experiment is that platforms continue to vary in their ability to counter manipulation of their services. Facebook-owned Instagram shows how this variation exists even within companies. Instagram remains much easier to manipulate than Facebook and appears to lack serious safeguards. Tellingly, the cost of manipulating Instagram is roughly one tenth of the cost of targeting Facebook.

In 2020, Twitter is still the industry leader in combating manipulation, but Facebook is rapidly closing the gap with impressive improvements. Instagram and YouTube are still struggling behind, but while Instagram is slowly moving in the right direction, YouTube seems to have given up. TikTok is the defenceless newcomer with much to learn.

Despite significant improvements by some, none of the five platforms is doing enough to prevent the manipulation of their services. Manipulation service providers are still winning.

This ongoing and evolving threat has underscored the need for a whole-of-society approach to defining acceptable online behaviour, and to developing the frameworks necessary to impose economic, diplomatic, or legal penalties potent enough to deter governments, organisations, and companies from breaking the norms of online behaviour.

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

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

> The evidence is clear: schemers around the world take every opportunity to manipulate social media platforms for a variety of commercial, criminal, and political reasons.

INTRODUCTION

One year ago, the NATO StratCom Centre of Excellence carried out a ground-breaking experiment to assess the ability of social media companies to counter the malicious use of their services. We showed that an entire industry had developed around the manipulation of social media[1], and concluded that social media companies were experiencing significant challenges in countering the manipulation of their platforms.

Since then the companies committed to improving their defences, especially ahead of the US 2020 presidential elections. Facebook,[2] Google,[3] and Twitter[4] have all updated their policies and increased transparency regarding the manipulation of their platforms during the past year, but it continues to be difficult, often impossible, to independently assess the effectiveness of their efforts.

In September 2020 the European commission presented an assessment of the effectiveness of the Code of Practice on Disinformation one year after implementation. The commission concludes that the social media platforms have put policies in place to counter the manipulation of their services, but lack of transparency is still too great to enable a thorough evaluation of the impact of social media manipulation.[5] As a response to this enduring challenge, the European commission's recently launched *Action Plan for Democracy* outlines a commitment by the Commission 'to overhaul the Code of Practice on Disinformation into a co-regulatory framework of obligations and accountability of online platforms'.[6]

As antagonists and opportunists continue to ramp up their efforts to manipulate social media, researchers, intelligence services, in-

dependent interest groups, and the social media companies themselves have published numerous reports over the past year detailing attempts to manipulate conversations on these platforms. The Ukrainian Security Services (SSU) documented their recent discovery of a number of advanced bot farms,[7,8,9] and many other agencies have reported on the continued activity of the now-notorious Internet Research Agency in Russia,[10] and on a multitude of state actors and other organisations ranging from Iran[11] and China[12] to lobby groups[13] and terror organisations[14] involved in social media manipulation. There are also many academic articles and studies detailing the systematic manipulation of conversations, including our own quarterly *Robotrolling*[15] report and several recent studies on the US presidential election.[16]

The evidence is clear: schemers around the world take every opportunity to manipulate social media platforms for a variety of commercial, criminal, and political reasons. It will remain 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 decided to re-run the research[17] we conducted in 2019 using experimental methods to assess the ability of social media companies to counter manipulation on their platforms. In order to further refine our methodology for this iteration of the experiment, we decided to focus on a smaller number of posts manipulated by a smaller number of manipulation service providers, which allowed us to track differences among the responses of the platforms studied, rather than the relative performance of the manipulators. Another change was the addition of the Chinese-owned social media platform TikTok to our experiment. TikTok has become one of the fastest growing social media platforms, currently ranked as the seventh largest with almost 700 million active users.[18]

In the context of the US presidential election we partnered with US Senators Chuck Grassley (Republican Party) and Chris Murphy (Democratic Party) to assess to what extent their social media accounts in particular, and verified social media accounts in general, are protected against manipulation.

Our experiment provides original, and much needed, insight into the ability of social media companies to counter the abuse of their platforms.

The Social Media Manipulation Industry

Many of the conclusions from our initial report, *The Black Market for Social Media Manipulation*,[19] and from last year's[20] iteration 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 and spoilers seeking to influence elections, polarise public opinion, sidetrack legitimate political discussions, and manipulate commercial interests online.

The 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 program. 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.[21] 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. They can cost several hundred dollars to purchase and often remain online for years.

Developments in 2020

During the past year, most indicators have signalled that manipulation service providers are prospering. They are upgrading their technical capabilities and improving the marketing of their services. The larger providers are updating their frontend and backend systems as well as adding new services. There is evidence to suggest that the volume of engagement for sale is increasing. We have found manipulation service providers offering up to 10 million fake views on Twitter and Facebook, 50 million fake views on Instagram, and as many as 100 million fake views on IGTV. There is also a greater plurality of services on offer with an increase in manually controlled audience-specific services. We have identified what we assess as being possible efforts to introduce AI text generation for fake comment delivery.

Despite increasing pushback from a number of social media platforms, all major manipulation providers identified by us have remained in business and several new actors have

emerged. Together, the three manipulation providers we used for this experiment claim to have completed more than 17 million orders; one of them with 27 employees serving more than 190 000 regular customers.

Social media manipulation continues to be cheap and readily available through a multitude of professional social manipulation service providers. In certain cases, simple automated manipulation is no longer available and only well-resourced manipulation service providers are able to manipulate all platforms. While we are seeing evidence that platform pushback is starting to have an effect, in 2020 manipulation service providers remain in the lead in the digital arms race.

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 the past year the manipulation industry had become increasingly global and interconnected. A European service provider will likely depend on Russian manipulation software and infrastructure providers who, in turn, will use contractors from Asia for much of the manual labour required. Social media manipulation is now a global industry with global implications.
3. The openness of this industry is striking. Rather than lurking 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 still advertise openly on major social media platforms and search engines.

> We spent 300 € and received 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.

THE EXPERIMENT

The aim of our experiment was twofold; first, we aimed to further develop and test a methodology for assessing the ability of social media platforms to identify and counter manipulation using commercial manipulation service providers; second, we used our updated methodology to evaluate the performance of the social media companies in 2020 in comparison to their performance in 2019 as assessed by our initial experiment.

For this year's iteration of the experiment we used three reliable Russian social media manipulation service providers to buy engagement on Facebook, Instagram, Twitter, YouTube and, for the first time, on TikTok. We purchased engagement with accounts set up specifically for this experiment. All posts in our purpose-made accounts were of an apolitical nature to avoid any risk of actual impact beyond the framework of the experiment.

In contrast to the 2019 experiment, we did not seek to assess the performance of manipulation service providers; instead we selected reliable providers and monitored their service delivery to encourage them to do their utmost to deliver the services we purchased. The 2020 experiment was designed primarily to test the ability of social media companies to withstand manipulation from well-resourced commercial manipulation service providers. Setting up the experiment this way allowed us to better compare the relative performance of the social media platforms tested.

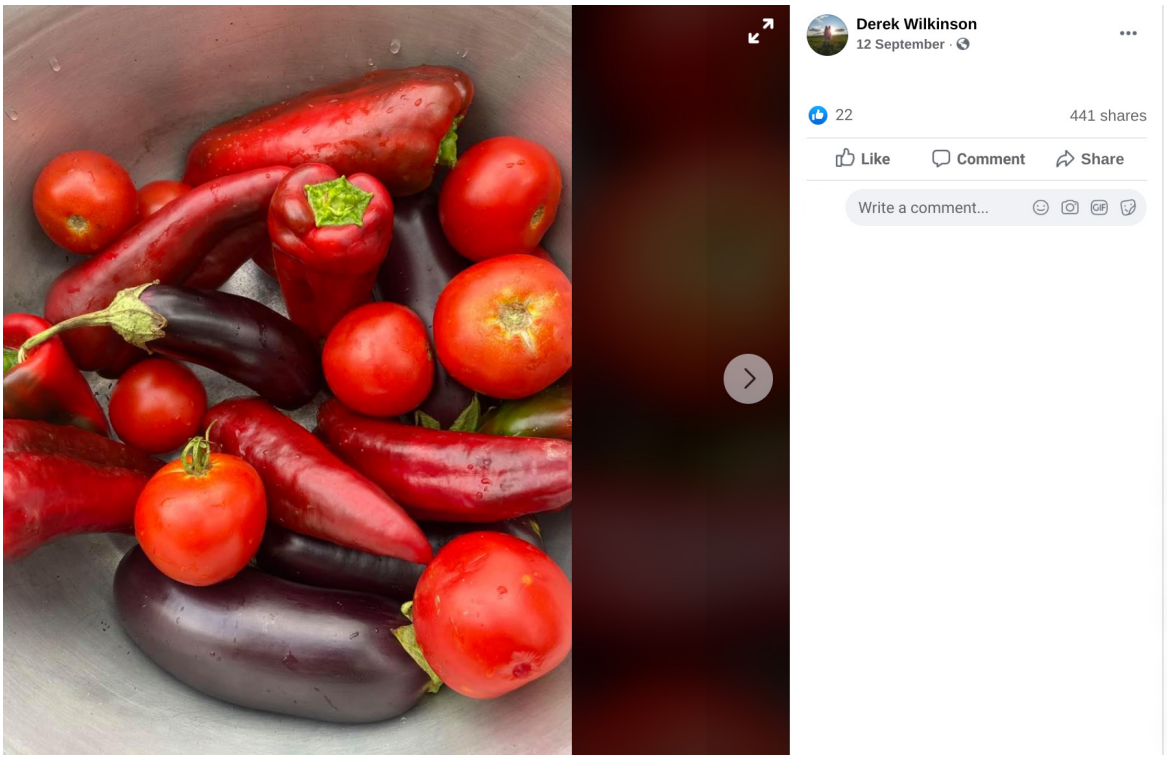
How is this relevant?

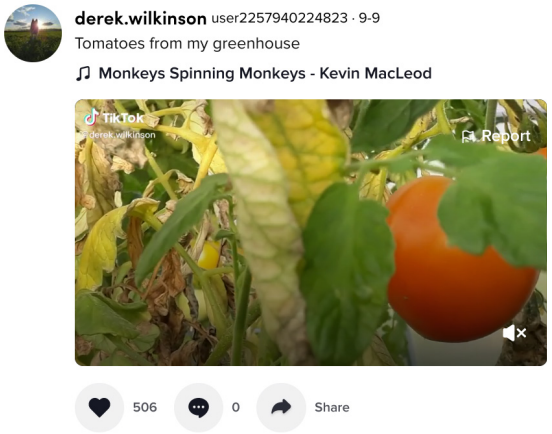
How is buying a thousand fake likes on a fake account relevant for the assessment of the overall ability of social media companies to counter manipulation? It could be argued that bought manipulation is more likely to be reported, and therefore detected, if it engages with current content that influences actual conversations online.

This time we were not testing the ability of social media managers or the public to detect and report inauthentic activity or the ability of the companies to remove posts based on the content and context alone. Instead we wanted to test the ability of the platforms to detect and remove networks of inauthentic users actively manipulating public conversations online.

By using commercial manipulation service providers and identifying and tracking the accounts they were using to deliver manipulation, we could judge the ability of social media companies to identify bot networks and inauthentic coordinated behaviour.

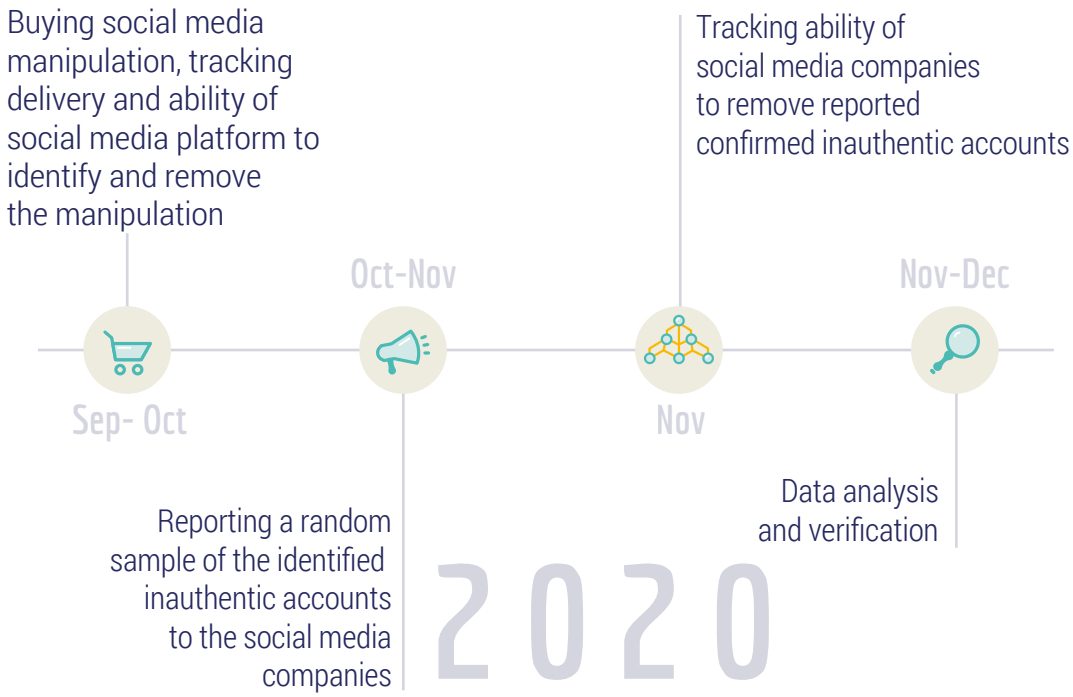
To lay to rest any potential difference between buying engagement on real and on fake accounts, we conducted a separate experiment to test if there is any difference in level of protection between verified accounts and ordinary accounts. For this experiment we partnered with US Senators Grassley (R) and Murphy (D) and bought manipulation on one apolitical neutral post on the Facebook, Twitter, and Instagram accounts of both men.





The experiment was conducted the week before the US Presidential election, during a time when social media companies were expected to be on special alert and to have implemented heightened detection mechanisms. This case study was intended to assess the maximum protection capabilities of the social media platforms studied.

One objection to our research design is that in selecting apolitical content, we may not have been testing platforms' ability to counter manipulation that might mislead platform users and impact the online conversation. However, conducting such an experiment would be hard to justify from an ethical perspective [see page 14]. In any case, the platforms claim to implement algorithms to detect platform misuse broadly, not just within the context of information operations. Facebook's community standards, for instance, prohibit specific behaviours (inauthentic behaviour, and especially coordinated inauthentic behaviour), rather than just specific content.[22]



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. Last year we spent 300 € to buy 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.

This year we spent 300 € and received 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. Below, we compare the cost of our basket of manipulation services to assess whether prices are rising year by year. On average, services were a bit more expensive than in 2019, but still roughly the same as in 2018.

Our work was carried out during six weeks in September and October 2020. 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 engagement to the social media companies and continued monitoring to measure the time it took for the platforms to react.

Five steps of the experiment

During the experiment, we recorded how quickly the manipulation service providers were able to deliver their services. We then collected data on how the five 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

An important part of this experiment was minimizing the risks involved, to ensure that private individuals were insulated from the experiment to the highest degree possible. While it would have been possible to design

an experiment to assess whether commercially purchased manipulation can influence public conversations, such research would be unethical in that it would interfere with genuine discussions and undermine important values such as freedom of speech.

Our experiment was set up so that we could minimize risk and carefully monitor any inadvertent effects. In order to achieve this, we chose to buy the fake engagement—views, likes, comments, and follows—using our own fake accounts. We continuously monitored our accounts to ensure there would be no authentic human engagement on them. We also chose apolitical and trivial content to engage with, and all bought engagements were strictly designed and monitored to minimize risk to real online conversations.

For the case study in which we engaged with content on the social media profiles of the two senators, we agreed with them in advance which posts would be targeted and how. We made sure to engage only with apolitical, dated content without any ongoing authentic conversations. We had also prepared interventions to alert any real users of our experiment had we noted any authentic interactions with the targeted posts.

Throughout the experiment we did not observe any indication that our false engagement had been noticed by authentic online users. 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 white hat programs of the social media companies themselves; these programs recognise the importance of external security researchers while emphasizing the importance of protecting the privacy, integrity, and security of users. We interacted with the two real, verified accounts only after having obtained explicit consent from the senators. 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.[23]

Finally, we made every effort to minimize the amount of bought engagement, to avoid unnecessarily supporting manipulation service providers. This year we reduced the number of posts engaged with and capped the amount spent at 300 €.

”Surprisingly 10 € will buy more than a thousand comments on Facebook-owned Instagram, but the same amount will only fetch 130 comments on Facebook.

ASSESSMENT OF THE PLATFORMS

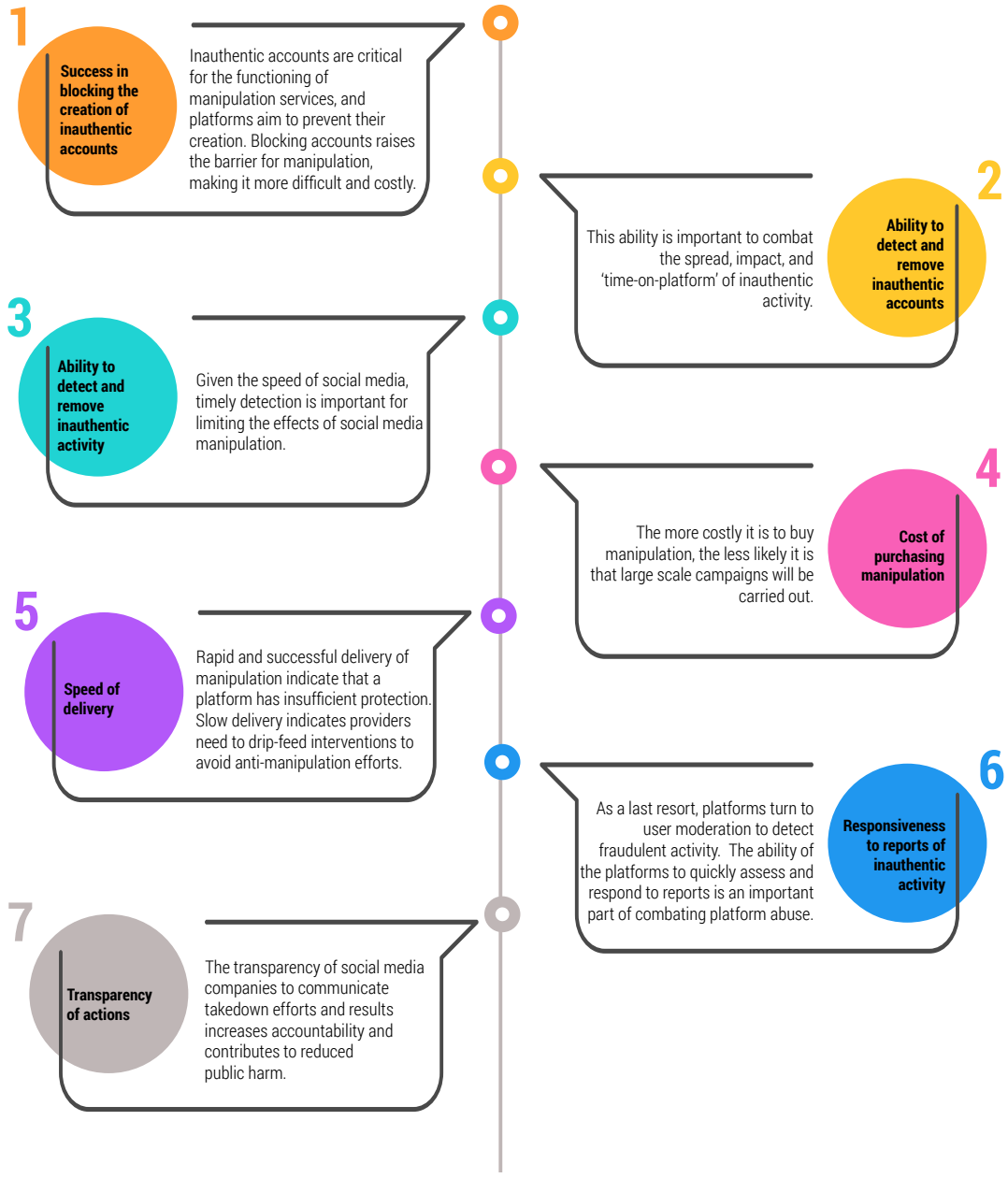
Our assessment criteria

We assessed the performance of the five 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, and 7) Transparency of efforts. We have further developed and refined our criteria since last year's experiment.

For this year's iteration we removed criteria assessing the ability of social media platforms to undo the activity of inauthentic accounts. While it is important that the activity of inauthentic accounts is removed together with the accounts themselves, this has become increasingly difficult to assess. Last year we determined that false likes bought on Instagram remained even after the account used to deliver them had been blocked.

The seventh criterion is new for this iteration. It was added to acknowledge the transparency of platform policy enforcement. We argue that increased transparency will reduce public harm by holding antagonists responsible and by improving users', researchers', and politicians' awareness of the scale and effects of social media manipulation.

These criteria can 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
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 indicate 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.

1. Blocking the creation of inauthentic accounts

Blocking the creation of fake or inauthentic accounts is perhaps the most important step social media platforms can take to prevent abuse. The easier and faster it is to create fake accounts, the easier it will be to manipulate social media activity.

In order to assess the ability of the social media companies to protect their users from manipulation, we attempted to create fake accounts using a variety of methods. We already knew that resellers of fake accounts provide detailed instructions on how to prevent them from being blocked; these typically include a set of preconfigured cookies to use with the account, dedicated proxies, specific IP locations that should be avoided, and other configuration descriptions required in order to run the account without getting banned. It was clear from these detailed instructions that significant improvements had been made in detecting suspicious account activity.

The biggest change this year was that Facebook had added a number of new features for blocking inauthentic accounts. It now requires significant effort to bypass Facebook's protection protocols; the account-generation industry must overcome these obstacles and some of the larger suppliers now provide specific tools and guidelines to help their customers.

Ultimately, however, none of the protection measures currently in place are robust enough to stop persistent users or organisations from creating inauthentic accounts on any of the platforms we studied. The continued low cost and effectiveness of manipulation services is proof of this, as we will see in the coming chapters.

YouTube, Instagram, and TikTok are in dire need of improved protections. This is especially surprising since Instagram and Facebook are owned by the same company, yet their protection protocols are worlds apart.

2. Removing inauthentic accounts

The longer bots and inauthentic accounts used to deliver 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.

Last year only 17 per cent of the inauthentic accounts identified on Instagram and YouTube had been removed after six weeks making them the worst-performing platforms for blocking inauthentic accounts. Facebook ranked third, having removed 21 per cent. The least-poorly-performing platform was Twitter, which succeeded in removing 35 per cent.

This year, the figures remained roughly the same at 0.4 per cent account removal per day. However, in calculating the half-life of all accounts identified as being used to deliver fake engagement, we can see that it has decreased from 128 days to 118 days, so there is actually a slight increase in the platforms' ability to remove them. The mean lifespan for inauthentic accounts has decreased from 224

days to 203 days before removal; although marginal, this is progress.

While social media companies continue to report that they block millions of inauthentic accounts annually, we are unable to verify the effectiveness of these efforts. On the contrary, inauthentic accounts seem to be active for quite a long time before they are detected and deleted. While it has become more difficult to create a fake account, we see only marginal improvements in decreasing the lifespan and half-life of recognised inauthentic accounts.

While their average lifespan of an inauthentic account on Facebook, Twitter, and Instagram[24] has decreased over the past year, Facebook stands out as the only platform with an increased half-life for inauthentic accounts. Twitter was the only platform that managed to significantly reduce half-life. This does not necessarily mean that the situation is worse on Facebook—these figures are calculated from small sample sizes, but they are sufficient to demonstrate there has been no measurable improvement.

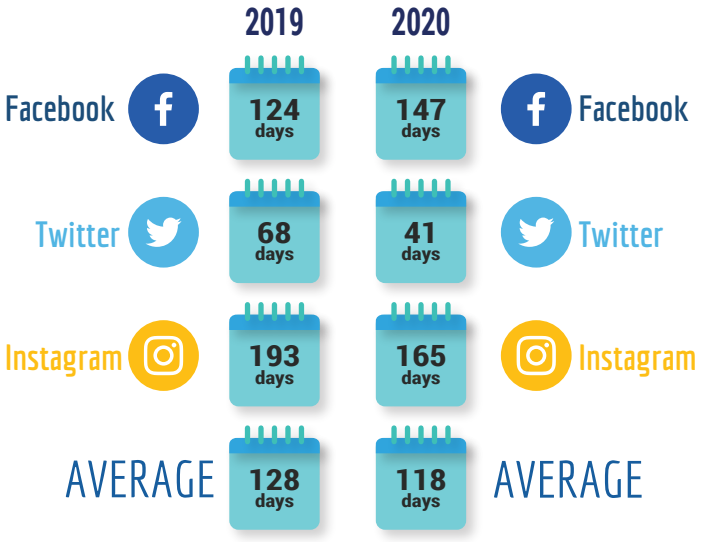
We conclude that there are significant differences among platforms, and that likely more can be done to reduce the half-life of active inauthentic accounts, especially with regard to the cheap and simple kind of inauthentic accounts that we tracked for this experiment.

3. Removing inauthentic activity

Removing inauthentic activity is the process of identifying fake engagement and removing it after it has been delivered. The faster the activity is removed, the smaller the effect it will have on social media conversations.

Last year we showed that social media companies struggled to identify and remove fake activity as the vast majority of all the fake engagement was still online four weeks after de-

BOT HALF LIFE BY PLATFORM



livery. The platforms performed relatively well only with regard to fake followers—roughly half had been removed after four weeks.

This year's results were roughly the same for Instagram and YouTube, with few engagements removed. Our new subject, TikTok, did equally poorly. Of a total 337 768 delivered fake engagements across all platforms, more than 98 per cent remained online and active after four weeks, and even discounting fake views, more than 80 per cent of the 14 566 fake engagements delivered remained active after a month. Twitter and Facebook showed progress this year, demonstrating active efforts to counter manipulation during the four-week test cycle.

Facebook's removal of inauthentic accounts resulted in a U-shaped pattern of the false engagement being added, removed, and then replaced by the manipulation service providers. Fake engagement gradually began to be removed after five days and reached maximum removal after about two weeks. On Twitter, some manipulation service providers had their engagement removed after roughly 12 hours. To bypass the protection protocols on Twitter and Facebook, the providers now drip-feed engagement; this seems to work rather well. It does, however, reduce the speed of manipulation service delivery, making it harder to effectively manipulate 'live' conversations.

All platforms still face significant challenges in countering fake video views. Only YouTube has a somewhat effective system for countering fake views; all other platforms continue to be practically defenceless. Fake views continue to be delivered quickly to Facebook, Instagram, Twitter, and TikTok and seem never to be removed. As this form of manipulation is the cheapest, we are led to speculate that the manipulation service providers might be using a different method to manipulate the number of views than they use for delivering likes and comments—one that manages to exploit flaws in how the platforms count views instead of relying on fake accounts to engage.

4. Cost of services

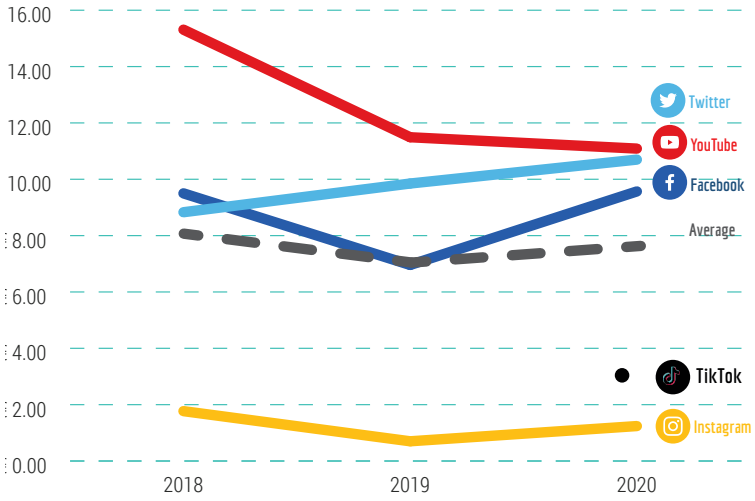
The cost of manipulation is a good indicator of how effectively social media platforms are combating manipulation. If accounts used to perform manipulation are removed, manipulation service providers have to spend time and money to replace them. When social media platforms redesign their service to make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

We compared the price of a basket of manipulation consisting of 100 likes, 100 comments, 100 followers and 1000 views from five Russian manipulation service providers to arrive at a mean price for 2020. The result is quite revealing. YouTube is the most expensive platform to manipulate at a cost of 11,12 € for the basket. Twitter joins YouTube at the high end with a price of 10,44 €. A Facebook manipulation basket is moderately priced at 8,41 €, while TikTok and Instagram are the cheapest at 2,73 € and 1,24 € respectively.

We found an increase of roughly 20 per cent in the price of a basket of manipulation services across all social media platforms from 2018 to 2020. However, a closer look at specific services reveals a tendency for comments to become more expensive, while the cost of likes remains the same or has got cheaper. Average price-levels are roughly the same as in 2018, meaning that there has been no significant change in prices over the past two years. There has been a slight decrease in the cost of manipulation services for YouTube, and a slight increase for Twitter. Prices remain roughly the same for Instagram and Facebook manipulation services as in 2018, however, prices for Facebook manipulation had dropped significantly in 2019.

Fake views continue to be the cheapest form of manipulation—about ten times cheaper than fake likes and follows, and roughly one hundred times cheaper than fake comments. While this ratio remains roughly the same, there is a significant difference between how much manipulation can be bought for 10 € on the various social media platforms. While 10 euro will buy 90 000 views on Instagram, it will only buy you 7000 views on YouTube. Surprisingly 10 € will buy more than a thousand comments on Facebook-owned Instagram, but the same amount will only fetch 130 comments on Facebook.

Over the past year, fake comments have become more expensive, which is a positive development, but social media manipulation in general continues to be cheap and readily available. There have been no significant price increases since last year and the differences in price between services and platforms remain relatively unchanged.



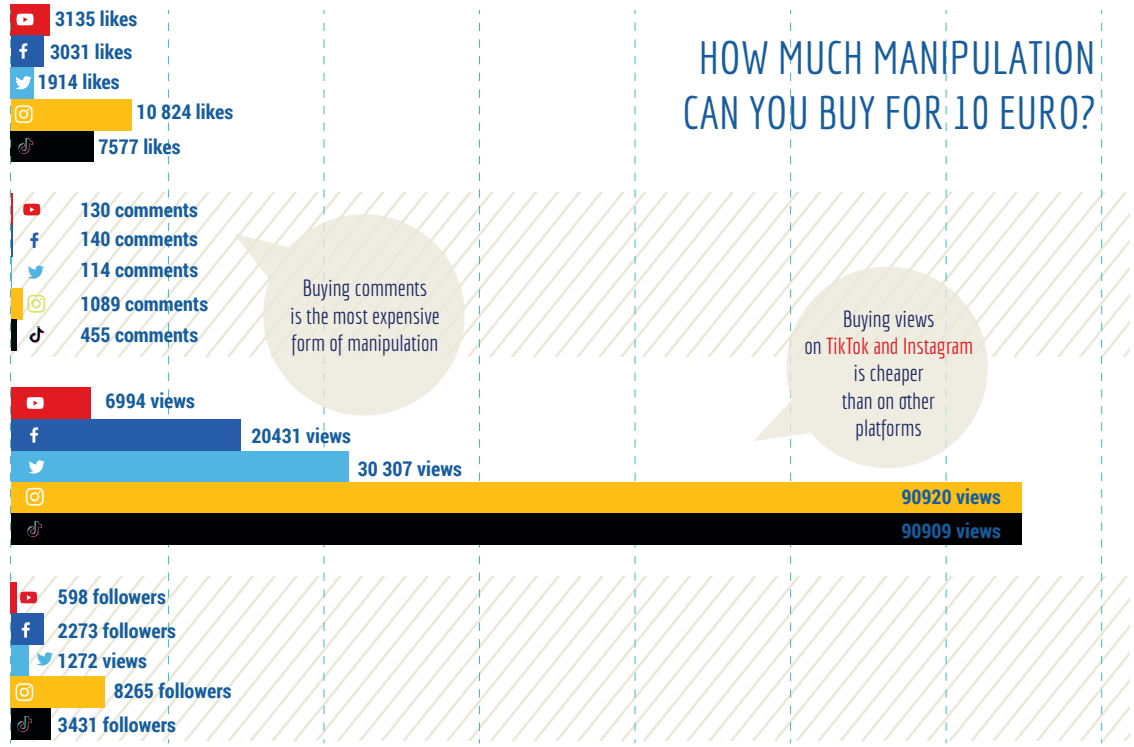
CHANGES IN THE COST OF A BASKET OF MANIPULATION SERVICES

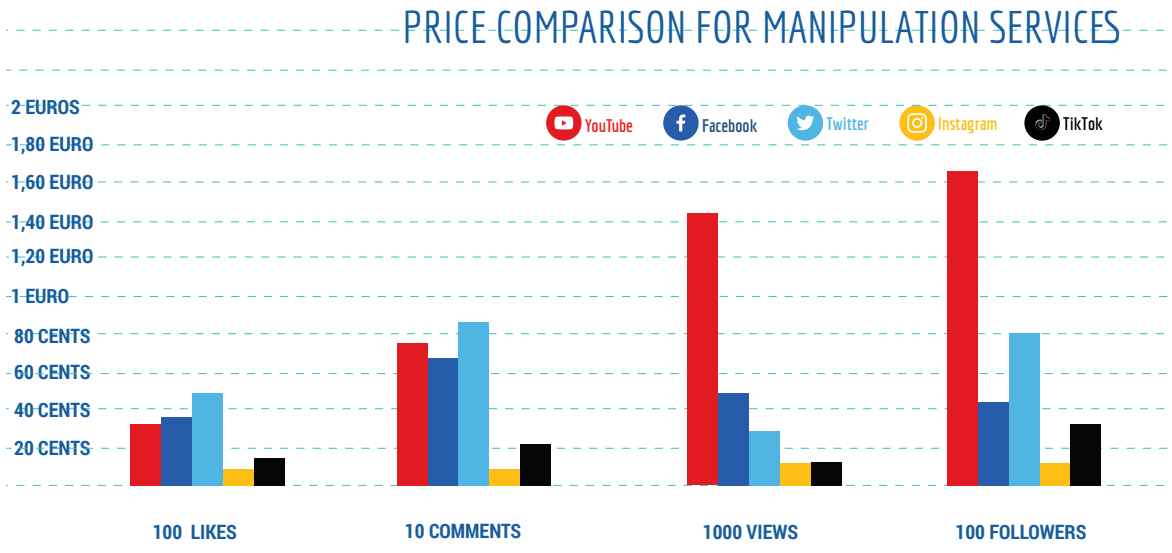
Facebook
100 likes
100 comments
1000 views
100 friends

Twitter
100 likes
100 comments
1000 views
100 followers

Instagram
100 likes
100 comments
1000 views
100 followers

YouTube
100 likes
1000 views
100 comments
100 subscribers


HOW MUCH MANIPULATION CAN YOU BUY FOR 10 EURO?
3135 likes
3031 likes
1914 likes
10 824 likes
7577 likes
130 comments
140 comments
114 comments
1089 comments
455 comments
Buying comments is the most expensive form of manipulation
Buying views on TikTok and Instagram is cheaper than on other platforms
6994 views
20431 views
30 307 views
90920 views
90909 views
598 followers
2273 followers
1272 views
8265 followers
3431 followers
YouTube
Facebook
Twitter
Instagram
TikTok


PRICE COMPARISON FOR MANIPULATION SERVICES
2 EUROS
1,80 EURO
1,60 EURO
1,40 EURO
1,20 EURO
1 EURO
80 CENTS
60 CENTS
40 CENTS
20 CENTS
YouTube
Facebook
Twitter
Instagram
TikTok
100 LIKES
10 COMMENTS
1000 VIEWS
100 FOLLOWERS

5. Speed and availability of manipulation

Last year we found that manipulation services for Instagram in general were the most reliable, while comment services for Facebook, Twitter, and YouTube were the least reliable. This has changed over the past year and there is enough evidence to conclude that platform efforts to improve security have had some effect. This is most notable on Facebook, where automated comment services are no longer widely available. Instead, manipulation service providers now must pay real humans to deliver fake comments. While this service has existed for quite some time, the fact that cheaper automated services are no longer available on Facebook is a testament to the effectiveness of Facebook's efforts.

Although it is important to recognise and encourage any improvements made by the platforms, social media manipulation continues to be widely accessible and professional. The providers we used are highly professional and customer oriented, offering discounts, money-back guarantees, and 24/7 support.

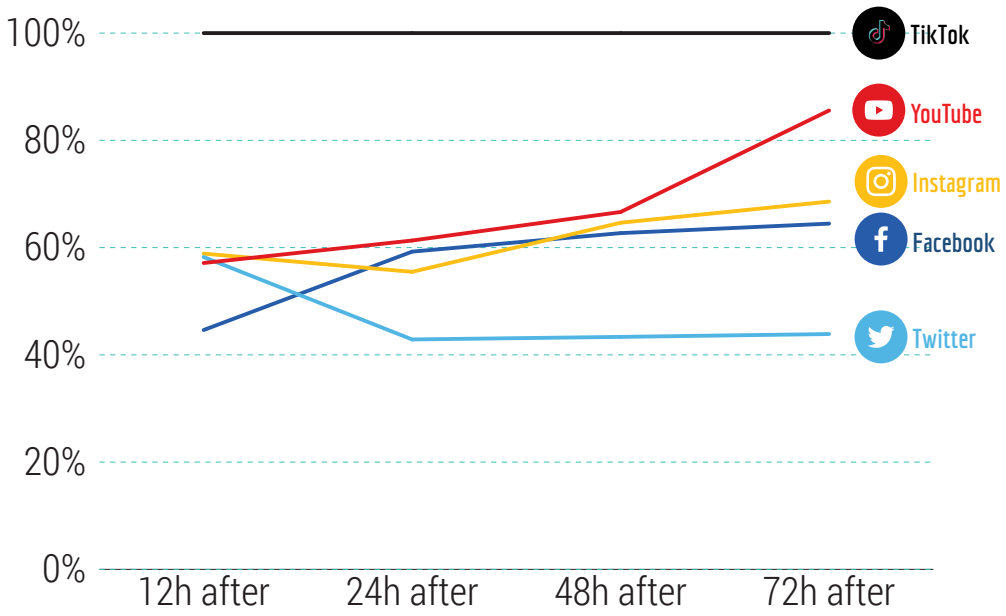
Platform pushback, primarily from Facebook and to some extent from Twitter, is challenging primarily automatic manipulation services, most notably comments. Another significant change is that comment services are now more often drip-fed, meaning that it now takes slightly longer for manipulation services to deliver purchased comments safely.

Comparing the speed of delivery of all manipulation services, excluding failed deliveries, across all social media companies studied, we found that roughly 60 percent of all manipulation services were delivered after 24 hours. Manipulations on Twitter arrive most quickly but are also removed most quickly (on average). TikTok performs worst; manipulations are delivered almost instantaneously and remain over time.

Views and likes are still delivered relatively quickly. Views are usually delivered within an hour, with the exception of YouTube, where it took as long as a month for all the fake views we had purchased to be delivered. On Instagram, we managed to get 250 000 fake views delivered within an hour. Likes were usually also delivered within an hour; when it took longer we surmise that this was likely because the manipulation service providers were slow to launch their services, rather than a result of platform countermeasures.

There has been an overall reduction in the speed of delivery of manipulation services. In 2019, most services were delivered in full within an hour. The main difference in 2020 is a reduction in the availability and speed of comment services on Facebook and Twitter. YouTube manipulation delivery is roughly equal to that of 2019; manipulation services are rather slow, and large orders of more than ten thousand views require days or weeks to be delivered in full. For the other platforms views are delivered as quickly and effectively this year as they were last year.

COMPARISON OF SPEED OF DELIVERY OF FAKE ENGAGEMENTS



6. Responsiveness

After the end of phase one of the experiment, we reported 100 random accounts used for social media manipulation to each of the platforms in our study, and then monitored how many accounts the platforms were able to remove within five days. It is worth reiterating that the accounts we reported were the accounts that delivered the manipulation we bought, meaning that we were 100 per cent certain the accounts we reported were engaging in social media manipulation. We reported these accounts anonymously.

In 2019, removal rates left a lot to be desired. Facebook removed 12 per cent, Twitter and Instagram 3 per cent, and YouTube 0 per cent of the reported accounts after three weeks. Given the speed of online conversations, this year we shortened the assessment time to five days.

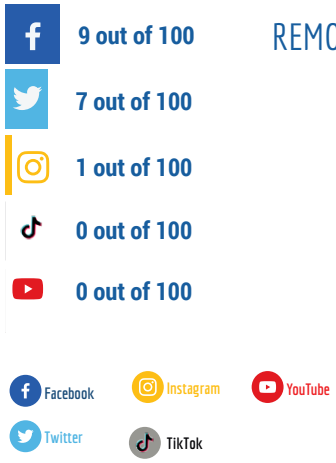
Our reporting produced better results in 2020, although the platforms' response is still woefully insufficient to combat manipulation effectively. Facebook was most successful, removing nine of the 100 accounts reported. Twitter followed close behind, removing seven accounts and temporarily suspending two others. Instagram removed just one of the reported accounts, and YouTube and TikTok didn't remove any.

To better compare this year's figures to last year's, we re-checked removals three weeks after the time of reporting. At that time Twitter was in first place, having removed twelve

accounts and suspended eight; Facebook had removed thirteen accounts, Instagram—three, TikTok—two, and YouTube still hadn't removed a single account.

This marginal increase suggests that time is not the factor that determines whether a reported account is removed. It was possibly the natural rate at which fake accounts get removed, rather than our reporting, that led to the small increases in removals over time.

It is clear that reporting an account as engaging in social media manipulation does not significantly contribute to any of the platforms removing or suspending that account, even if the validity of the report is confirmed. Indeed, in all cases where we received feedback about an account we had reported, it was the platform letting us know that the user in question had not violated the platform's terms of service. We suspect that platforms are using quantitative indicators rather than a manual assessment of reported accounts to determine abuse. Manual vetting of the reported accounts would have clearly demonstrated that, given the frequency and diversity of the engagement, the reported accounts were delivering paid services.



We conclude that reporting and moderation mechanisms must be improved so that a larger share of inauthentic accounts reported are removed, even if they are reported by a single user. It is not satisfactory that reported inauthentic accounts regularly escape sanction.

7. Transparency of efforts

All the major social media platforms studied for this report have increased their communication efforts regarding measures undertaken to counter abuse. Their efforts are in alignment with the European Code of Practice on Disinformation and the political pressure the platforms were facing ahead of the US presidential election of 2020. Today TikTok,[25] Twitter,[26] Facebook,[27] and Google[28] all publish transparency reports of various kinds, some of which relate to fake accounts and other forms of inauthentic coordinated behaviour.

Facebook reported that in Q2 2020, 1,5 billion fake accounts were removed for violating the community standards, a slight reduction from 2019, which was attributed to the platform's increased ability to block the creation of fake accounts.[29] Twitter's most recent report on fake account removal is from Q3/Q4 2019 when the platform reported that it had sanctioned 2,3 million accounts and suspended close to 800 thousand; Twitter also reported a 30 per cent increase in accounts sanctioned for platform manipulation.[30] We have been unable to find detailed information on fake

account removal or on platform manipulation in general for Google and TikTok. Google does provide information on advanced threats through its Threat Analysis Group.[31]

Twitter's safety blog,[32] Facebook's news archives,[33] and TikTok's Newsroom reporting on safety-related matters[34] provide further information pertaining to specific threats and manipulation activities, often focusing on specific content being taken down.

Facebook should be commended for the detailed takedown reports it is now publishing on a monthly and case-by-case basis. Twitter should be commended for publishing data that outside researchers can study. The continuous improvements by the Google Threat Analysis Group is also a positive step, even if the content of its reports has yet to reach the level of detail and transparency that Facebook is delivering.

While all of these efforts represent steps in the right direction, much more can be done in terms of contextualizing information, conducting more thorough audits and publishing the results, developing disclosure frameworks and collaborative transparency best practices jointly with other platforms, etc. From the information currently being provided by the platforms, it continues to be difficult to assess how the threat landscape is developing and how well counter efforts are working.

”Twitter remains the industry leader in 2020 but Facebook is rapidly closing the gap with its impressive improvements. Instagram and YouTube are still struggling behind, but while Instagram is slowly moving in the right direction, YouTube has made very few improvements. TikTok is the defenceless newcomer with much to learn.

ASSESSMENT OF EACH PLATFORM'S RELATIVE STRENGTHS AND WEAKNESSES

The performance of each social media platform is assessed based on the criteria introduced above. To assess the relative performance for each of the social media platforms we have given a rating to each company for each category. These qualitative ratings are based on a joint assessment by the research team.

Facebook

Facebook is the most effective social media platform at blocking the creation of inauthentic accounts. It uses advanced analytics and techniques such as requesting the users to upload videos to verify their accounts to identify them and keep fake accounts from making it onto the platform. In order to create a fake account on Facebook, advanced IP-address and cookie control is now necessary; this has prompted manipulation service providers to offer lengthy tutorials about how to prevent a fake account from being blocked. We conclude that creating fake accounts on Facebook has become significantly more difficult in the last year.

Despite being the most effective at blocking the creation of fake accounts, Facebook was the only platform studied that showed an increase in the half-life of active inauthentic accounts—currently an average 147 days. In this regard Facebook is significantly behind Twitter.

However, Facebook has made important strides in removing inauthentic activity. A portion of the likes we had purchased were gradually removed over a period of two weeks. We also observed a reduction in the speed of manipulation delivery, especially for fake comments, indicating that manipulation service providers can no longer automate such delivery.

The price of Facebook manipulation decreased between 2018 and 2019, only to return to 2018 levels again in 2020. This suggests that an increase in protection protocols is starting to translate into higher manipulation costs.

Facebook has not improved its ability to remove reported inauthentic accounts—at 9 per cent after five days, the ratio is still far too low to make an impact. Facebook did, however, increase its transparency and is currently publishing regular updates on takedown activities. On 19 November 2020, Facebook further refined and updated its community standards page,[35] and provided additional information regarding fake accounts.[36] We are still missing important details about how Facebook compiles its statistics and would welcome independent assessment and auditing of the data and reports published by the company. Facebook's current research initiatives are a crucial step in the right direction, and we look forward to studying the results.[37]

Facebook has had a good year. We have noted improvements in every category of measurement, although some are only marginal. We commend the platform's improvements in blocking fake account creation and removing fake activity, and its success in stopping some forms of automatic manipulation. However, the fake accounts that are created still survive too long, and Facebook's responsiveness to reporting leaves much to be desired.

Instagram

Even though Instagram is owned by Facebook, it is far less effective at countering platform abuse. In fact, Instagram is outperformed by its parent company in every category we assess. Manipulating Instagram remains cheap and effective.

We observed no significant improvements in blocking or removing fake accounts. However, manipulation delivery is slower now, and we saw that manipulation providers encountered challenges resulting in only partial deliveries for specific kinds of manipulation, such as mentions. It remains to be seen whether these 'wins' for the platform are temporary.

Instagram's failure to counter manipulation is perhaps most evident in the cost of manipulation. Instagram continues to be the cheapest platform to manipulate, even cheaper than TikTok. Moreover, prices are currently falling, suggesting that manipulation services are highly available.

In assessing Instagram's performance in comparison to 2019, we observed only marginal improvements—primarily that delivery of fake likes and comments are now delivered significantly slower than last year. Fake views are still delivered instantaneously.

We conclude that Instagram remains very easy to manipulate and whatever slight improvements it has introduced in the last year have made little difference. Instagram's parent company, Facebook, should share its growing expertise.

Twitter

Twitter continues to be the most effective platform overall at countering manipulation, outperforming Facebook in several areas but falling behind in others. Facebook is slowly gaining on Twitter's lead.

We observed less pushback against the creation of fake accounts by Twitter this year than in 2019. None of our own accounts were blocked this year, and we observed fewer counter measures in 2020 than in 2019. Facebook's accomplishments show that more can be done by Twitter in this field.

We observed continued improvements by Twitter in all other categories. Most significantly Twitter stands out for its ability to remove inauthentic accounts. Fake accounts are disappearing 40 per cent faster than in 2019, making Twitter three times faster than Facebook.

Twitter has also improved its ability to remove fake activity and to slow down the speed of delivery of inauthentic engagement. Although half of the bought likes were delivered within an hour, Twitter had effectively removed many within 12 hours. After 24 hours 80 per cent had been removed and after two days virtually all of the fake likes had vanished. However, the drip-fed engagement proved more difficult. Twitter blocked a significant proportion (as much as 80 per cent), but the remaining 20 per cent remained four weeks after delivery.

Of the 100 random accounts we reported to Twitter, 7 per cent were removed and 2 per cent were suspended after five days. This is an improvement over last year, when only 3 per cent were blocked three weeks after reporting, but the overall performance is still far too low to make an actual impact.

Twitter was the only platform to cause a year-by-year increase in the price of manipulation services, however, the rate of increase slowed from 15 percent between 2018 and 2019 to only 4,5 per cent between 2019 and 2020.

In terms of transparency Twitter lags behind Facebook in updates and regular takedown disclosures but excels in making datasets available for external review by the scientific community. Independent audits of methodology would be useful for Twitter and for the other platforms.

Twitter improved in all categories except for blocking account creation and should be commended for the strides it has made in its ability to remove inauthentic accounts and activity.

We conclude that Twitter maintains its position at the most effective platform at countering manipulation, but it can do more to block the creation of fake accounts and to increase its ability to remove reported inauthentic accounts.

YouTube

YouTube remains mired in its position as the least effective of the four major platforms. When we added TikTok this year, YouTube lost the distinction of being the worst platform, but only because TikTok is worse.

We are unable to accurately assess the half-life of identified but active YouTube accounts used to deliver manipulation because of the platform's lack of transparency; it simply isn't possible for outside researchers to identify the accounts delivering the most inauthentic content. There was virtually no reduction in fake comments or the accounts used to de-

liver them; only three of 380 fake comments delivered were eventually removed. On the basis of this performance, we conclude that YouTube removed few or no inauthentic accounts on its platform. Furthermore, YouTube removed none of the accounts we reported for inauthentic activity, a result worse than last year.

YouTube does a better job than the other platforms in removing inauthentic views, as none of the other platforms managed to remove any of the inauthentic views we bought. We observed a ten per cent reduction after three to four weeks, and a 50 per cent reduction after eight weeks. However, we observed that fake views were removed for some videos and not for others, indicating an uneven distribution of YouTube counter-measures. Interestingly, we experienced significant over-deliveries this year; manipulation service providers delivered 70–90 per cent more engagement for some YouTube services than we had bought.

The delivery of inauthentic likes is nearly instantaneous on YouTube, while inauthentic views and comments are either drip-fed or added in increments. The average speed of delivery, across all services, for YouTube was equal to that for Instagram and Facebook after 12 hours, but after 72 hours we observed 85 per cent delivery on YouTube compared to only 45–60 per cent for the other platforms. This means that manipulation services, on average, can be delivered to YouTube much more quickly than to Twitter, Instagram, or Facebook. But, as mentioned above, YouTube is the only platform able to counter fake views. Currently all the manipulation providers we tested drip feed fake views onto YouTube at a rate of 500 to 4000 views per day, with significant fluctuations over time. It took the manipulation providers 6–12 hours to deliver the first 1000 views, and then another four weeks to reach 10 000 views. YouTube outperforms every other platform tested in this category but isn't able to stop fake views from appearing on its platform.

YouTube maintains its position as the most expensive platform to manipulate, however it was the only platform for which the price of manipulation services had decreased since 2019. Although the 3,5 per cent reduction is smaller than the 25 per cent reduction from the year before, the trend is still negative; YouTube is getting cheaper to manipulate over time and fewer inauthentic accounts and engagements were removed in 2020 than in 2019. The primary positive development was a continued, and somewhat enhanced, ability to slow down the delivery of fake views.

Although YouTube has made some advance in transparency this year, it still demonstrates a significant lack of transparency, providing much less in the way of facts, statistics, and takedown reports than Facebook and Twitter.

It is clear that YouTube needs to prioritize combatting inauthentic behaviour. It is embarrassing that several large manipulation service providers maintain accounts on YouTube that publish promotional material and how-to videos for manipulating the platform. Furthermore, manipulation service providers continue to advertise their services successfully through Google Ads—YouTube's parent company.

TikTok

The story about TikTok's ability to counter manipulation is a sad one, as manipulating TikTok is cheap, fast, and reliable. TikTok seems unable to identify and remove even the cheapest forms of manipulation from its platform and provides its users with only the most basic protections.

For the duration of our experiment, TikTok did not manage to remove any of our fake accounts or any of the fake engagements we had purchased. Engagements were delivered almost instantaneously; only comments required between 6 and 12 hours to be delivered, indicating that these may have been manually delivered onto the platform.

Given the ease and availability of manipulation, we were surprised to discover that manipulation services for TikTok are about 45 per cent more expensive than for Instagram on average (manipulating TikTok in turn is one tenth the price of manipulating Twitter on average). While TikTok comments are about twice the price of Instagram comments—roughly 2 EUR for 100 fake comments—fake TikTok views are as cheap as fake Instagram views, with a median price of 0,11 EUR for 1000 inauthentic views.

TikTok's countermeasures were largely non-existent. For the categories manipulation costs and speed we gave the platform the benefit of the doubt, noting some small possible responses. And on a positive note, TikTok has started a program to disclose information about its efforts to counter abuse on the platform.

We conclude that TikTok is far too easy to manipulate—so easy that perhaps the best way to describe the platform is simply as undefended. The only positive aspect of TikTok's ability to counter manipulation is that it casts the efforts of the other social media platforms in a much better light

Relative performance.

The most important insight from this iteration of our study is that there continue to be significant differences among platforms in their ability to counter the manipulation of their services. From the near defenceless TikTok to industry leader Twitter, all platforms perform differently in different categories.

Twitter remains the industry leader in 2020 but Facebook is rapidly closing the gap with its impressive improvements. Instagram and YouTube are still struggling behind, but while Instagram is slowly moving in the right direction, YouTube has made very few improvements. TikTok is the defenceless newcomer with much to learn.

Despite notable improvements by some, none of the five platforms we studied are doing enough to prevent the manipulation of their services. The manipulation service providers are still winning the digital arms-race.

ASSESSMENT OF THE PLATFORMS RELATIVE STRENGTHS AND WEAKNESSES

**In red - relative change from 2019*

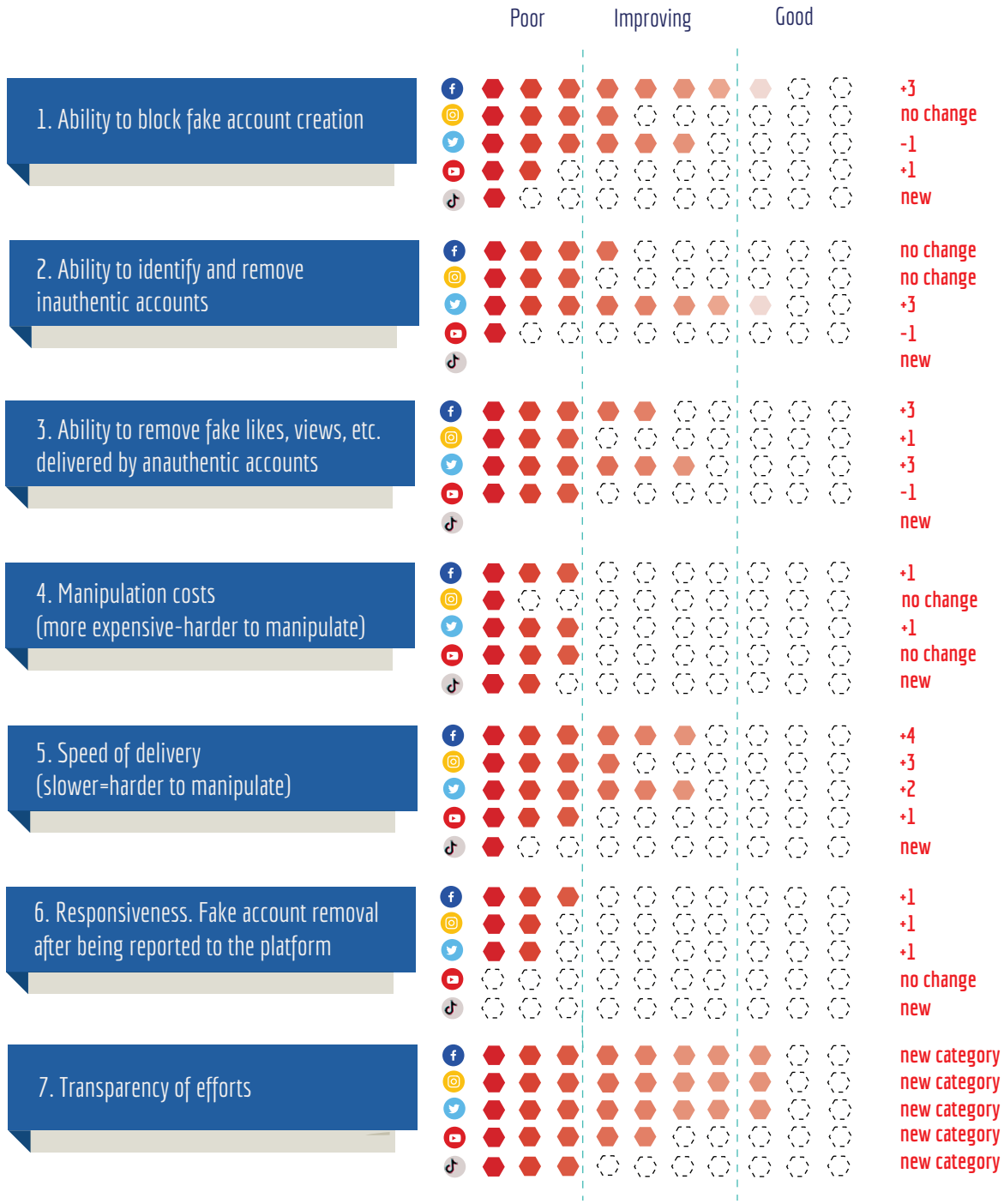


Illustration of relative performance of Twitter (1st), Facebook (2nd), Instagram (3rd), Youtube (4th) and TikTok (5th). Manipulation service providers are still winning.

[x] Errata. A previous version of this report showed the wrong relative figures by mistake

[xx] For this year we have included a new category "Transparency of effort". We have also removed the category assessing the ability to "undo historic activity"

What other content were the bots targeting?

Because of changes in the way in which social media platforms share data and allow for external data access, it is challenging to identify, monitor, and analyse which other types of accounts and content were targeted by the commercial bot networks we studied. On most of the platforms, we were unable to identify which accounts were being used to deliver views and likes; even when we could identify an account as being responsible for specific inauthentic efforts, we are often unable to see the other activities of that account.

This section provides some insight into the activity of manipulation service accounts by looking at the activity of accounts identified on Twitter, as Twitter alone provides the necessary transparency to allow us to discover which other coordinated activities the accounts engaged in.

We identified a network of inauthentic accounts that promoted 71 different topics over the course of a week, including financial services, politics, health, pornography, gambling, sports, and online influencers.

The financial services being promoted consisted mostly of spam-like cryptocurrency ads, while the goods and services promoted included chili sauce, makeup artists, cyber security specialists, and a monument installation service in St. Petersburg.

We identified bots promoting material relating to the US election. These efforts appeared to be bipartisan; we identified accounts promoting both pro-Biden and pro-Trump content. Other accounts pushed conspiracy material surrounding the Catalan independence struggle and the poisoning of Aleksej Navalny. We also identified accounts discussing the politics of Mexico, Brazil, and Ukraine, calling out alleged anti-Semitism in Lithuania, and expressing xenophobic content.

As noted in the 2019 report, influencers interested in inflating their follows continue to be an important source of business for the commercial manipulation bot industry. This year we observed that sports and entertainment influencers had joined the usual fashion influencers using these services. Perhaps the most surprising observation was that bots were being used to inflate engagement on a brand belonging to a major American artist.

Sadly, we also observed attempts to spread disinformation surrounding the COVID-19 pandemic, although with little apparent success as the posts targeted contained fringe content that attracted few or no organic engagements.

Our conclusion from 2019 stands: manipulation services are still being used primarily for commercial purposes, but political actors are making minor forays into manipulating public discourse.

CASE STUDY: US PRESIDENTIAL ELECTIONS 2020

Are the platforms protecting verified accounts during elections?

manipulation such as the artificial inflation of likes, views, and comments.[38, 39, 40] In light of the numerous recent attempts by hostile actors to interfere in democratic elections by weaponizing social media platforms, this is important and commendable work. Furthermore, it is vital that the work of the social media companies be independently assessed and audited to ensure that their countermeasures are robust and effective.

This experiment was conducted on the cusp of the US presidential election to assess the general ability of social media platforms to counter manipulation. In order to discover whether verified accounts on social media platforms are better protected than ordinary accounts, particularly during a US presidential election, we partnered with US Senators Grassley (R) and Murphy (D) to conduct a unique experiment.

Using commercial social media manipulation service providers, we bought fake shares, likes, and comments for two social media posts—one from each senator—on Twitter, Facebook, and Instagram. These platforms were chosen because the two senators are active on these platforms and their accounts contain content that met the stringent criteria we developed to reduce the risk of inadvertently influencing any current political discussions.

We chose six neutral, apolitical posts for our case study—one from each senator on each of the three platforms. The posts were at least one month old and didn't have any active engagement, such as comments or views. We then purchased the smallest baskets of manipulation services available: 1000 likes, 50 comments, and 50 shares on Facebook; 1000 likes and 50 comments on Instagram; and 1000 likes, 50 retweets, and 50 replies on Twitter.

For Facebook the comments were deployed as tasks, meaning that we gave instructions to the manipulation service provider, who in turn passed them on to the people performing the tasks. A higher degree of automation is possible on Twitter and Instagram, so here we provided pre-written comments that were posted by bots. Pre-written comments have the added benefit of being easier to trace after they are delivered.

After having purchased the fake engagements we monitored the posts closely for eight days before reporting a random sample of the accounts observed delivering the fake engagement. We then continued to monitor the posts for another five days to observe the proportion of reported accounts removed by the platforms and assess the enforcement capabilities of the social media companies.

Instagram

We paid 7,3 USD for 1803 likes and 103 fake comments delivered successfully on Instagram.

The comments we bought on Instagram were delivered to the senators' posts within one hour and remained in place throughout the experiment; the bought likes started appearing after an hour. Ninety per cent of the inauthentic engagement we purchased was delivered

within the eight-day observation period. The likes and comments were all delivered by relatively simple fake accounts that had few or no friends and no original content.

Facebook

We paid 26,3 USD for 2205 likes, 40 comments, and 110 shares delivered successfully on Facebook.

All of the fake likes were delivered within one hour, but over the subsequent hours and days the number of likes steadily declined. The manipulation service provider added more likes after 4 days to compensate for likes that had been removed, by which time roughly half of the original fake likes had been removed. At the end of the eight-day observation period roughly half the number of fake likes we had purchased remained.

The fake shares were delivered over a 24-hour period and at the end of the observation period all the fake shares remained without any reductions. The manipulation provider encountered some challenges delivering the fake comments as only 40 of 100 comments were delivered by the end of the observation period. All 40 comments were delivered between 6 and 24 hours after purchase.

We determined that most of the accounts delivering the inauthentic engagement were genuine accounts/users getting paid for delivering the engagement. However, some of the accounts looked suspiciously like simple bot accounts with little content and few friends. Almost all of the accounts seemed to originate in Asia. We determined this by looking at the names and languages used in other posts made by the accounts.

Twitter

We paid 28.4 USD for 220 likes, 75 replies, and 95 retweets delivered successfully on Twitter.

For one of the accounts on Twitter, the likes were never delivered. For the other post, only about 20 percent of the likes had been delivered after the eight day observation period. However, at the end of the experiment roughly half of the likes had been delivered, which suggests that the likes were being drip-fed slowly to avoid being blocked by Twitter.

The retweets we purchased were delivered successfully to both posts. Numbers peaked within about six hours, after which a few of the retweets began to disappear. After eight days about ten percent of the retweets had disappeared. Seventy-five per cent of the inauthentic replies were delivered within 24 hours and none had been removed by the end of the experiment.

Analysis

Instagram was the easiest of the three platforms to manipulate. Manipulation services were delivered rapidly and nearly in full. Delivery of Facebook services was staggered; those that were delivered all in one go tended to disappear quickly as the platform's anti-manipulation algorithm rooted out the fake activity. Apparently, the manipulation service provider delivered the inauthentic likes slowly to circumvent prevention mechanisms.

The manipulation service provider we used did not offer automatically generated comments or comments delivered through automated accounts for Facebook; instead the current model is to pay real people to perform specific tasks. Manipulation of this type is harder to detect, but also harder to scale.

This suggests that Facebook is winning against the manipulation providers in this area by denying them the ability to deliver fake comments automatically.

We observed no difference between the protection offered to senators and that provided to regular users on any of the three platforms in respect of inauthentic engagements delivered and removed. The speed and completeness of manipulation services was broadly the same for verified and regular accounts. However, we observed a significant difference in the way reported accounts were handled. At the end of the eight-day observation period we reported a sample of the accounts identified as delivering inauthentic engagement and then measured how many had been removed five days after reporting. Instagram removed 16 of 40 reported accounts (40%), Facebook removed 10 of 60 reported accounts (17%), and Twitter removed 5 of 50 reported accounts (10%). Rechecking removal rates after three weeks, we found that Facebook hadn't removed any more accounts, but both Instagram and Twitter had roughly doubled removals, by 70% and 24% respectively. In removing reported accounts for the two senators, Facebook and Instagram outperformed their results for regular users significantly, while Twitter's performance remained constant. This suggests that reported accounts engaging with US political accounts may undergo additional scrutiny. Even so, given the simplicity of the reported accounts, manual assessment should have resulted in a much larger percentage of removals. Consequently, we are unable to rule out the possibility that reasons other than platform vigilance account for the difference.

Case study conclusions

A significant proportion of the inauthentic engagements we purchased was successfully delivered by the manipulation service provider. With some important exceptions, a large proportion of the delivered engagements remained undetected for the duration of the experiment. The accounts delivering the manipulation also remained active throughout the experiment, and we observed that these accounts continued to deliver purchased engagements for other commercial clients throughout the monitoring period.

The results of this small study on verified accounts are comparable to our general findings for 2020. There is no clear evidence to suggest that the platforms added any safeguards for verified political accounts to counter this form of manipulation during the current US election cycle.

Despite the improved removal of reported accounts, the results of this case study are similar to those of last year's case study when we bought engagement on posts by EU commissioners Dombrovskis, Jourová, Katainen, and Vestager. This indicates that the platforms have not made any major changes in how they protect verified accounts.

This leads us to conclude that verified institutional accounts are still no better protected against manipulation than any other accounts on the social media platforms. Our experiment also shows that at the height of the US presidential election it remained possible to buy fake engagements on the social media accounts of two US senators from a Russian manipulation service provider with relatively little pushback from the platforms in question. Efforts to protect official accounts and to counter bot networks and commercial manipulation must be significantly improved to protect online conversations.

ChuckGrassley
@ChuckGrassley

Happy natl chocolate chip cookie day. I like my cookies w a cup of coffee @ Hart Senate Office Building instagram.com/p/CDelBWOjBaT/…

8:40 PM · Aug 4, 2020 · Instagram

74 Retweets **16** Quote Tweets **489** Likes

Christina Gamble @Christinagablea · Oct 25
Replying to @ChuckGrassley
Chocolate chip cookies are such a great cookie for coffee, breakfast or just an afternoon snack

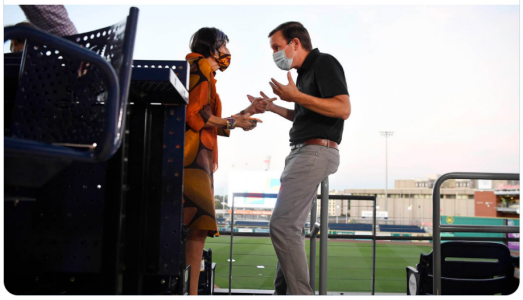
Marie Baird @MarieBaairrd · Oct 25
Replying to @ChuckGrassley
One day chocolate chip cookie will rise up and live out the true meaning of its creed

Kristen Morrow @Kristenmorrows · Oct 25
Replying to @ChuckGrassley
Mmmm...Chocolate chip cookie is my true love

Chris Murphy
@ChrisMurphyCT

Rosa and I arguing over the best pizza in New Haven.

> **Jessica Hill** @NutmegNews · Aug 21
> .@Rosa_DeLauro and @ChrisMurphyCT at @GoYardGoats for the #DNC2020 watch party



4:02 AM · Aug 21, 2020 · Twitter for iPhone

88 Retweets **19** Quote Tweets **989** Likes

Senator Chris Murphy
15 September ·

This will make you smile.

WASHINGTONPOST.COM
With its 4-year-old starring, a Connecticut family recreates famous movie scenes

571 104 comments 71 shares

> It is still far too easy to manipulate social media platforms, even at the height of a US presidential election.

CONCLUSIONS

Last year an important finding of our experiment was that the different platforms weren't equally bad—in fact, some were significantly better at identifying and removing manipulative accounts and activities than others. This year's iteration has made the differences even clearer as we have seen great improvements by some platforms, while others have reached new lows. **Investment, resources, and determination make a significant difference in the ability of social media companies to counter manipulation.**

It has also become evident to us that social media companies have different strengths and weaknesses. This seems to indicate that they don't share information about lessons learned and best practices amongst themselves. Many of the challenges they face could be addressed more effectively if the companies themselves 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. Telecom companies, online payment providers, web hosting services, search engines, and online advertising companies all need to come together to combat the digital manipulation industry.

It is still far too easy to manipulate social media platforms, even at the height of a US presidential election. Although we studied only commercial manipulation service providers, there is no doubt that other more maliciously inclined actors could use the same techniques for political or security-related manipulations. Given the number of takedowns and ample evidence of state-driven social media manipulation over the course of the past year, published and acknowledged by the platforms themselves, it is clear that antagonists continue to exploit social media loopholes to manipulate public discussions.

As antagonists find that the big social media platforms, primarily Twitter and Facebook, are slowly closing the door on them, it is reasonable to expect that they will look for other avenues of access, probably seeking to exploit and manipulate less defended platforms and online services. This risk underscores the continuing need to regulate digital services to ensure equal protections for online conversations regardless of where they occur.

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

More transparency is needed to understand the scope and effect of available manipulation services. In particular, more detailed information is needed regarding the "actors, vectors, targets, content, delivery mechanisms and propagation patterns of messages intended to manipulate public opinion".[41] 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 that allows watchdog agencies to compare reports from the different social media companies. Tech companies also need to be encouraged or forced to share technical data that would enable joint development of best practices and optimize 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 need to regulate the market for social media manipulation.

Political manipulation by domestic groups and foreign governments needs to be exposed and perpetrators must be held accountable. Violators can be deterred by economic, diplomatic, and criminal penalties. 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 social media companies over the past year, it is important that we continue to pressure them to do more to counter platform manipulation. Manipulation service providers continue to advertise and promote their services on the very platforms that they seek to undermine. It is embarrassing for the social media companies that they 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.

4. A whole-of-industry solution is needed

Social media companies will not be able to combat social media manipulation as long as there isn't a whole-of-industry solution to the problem. Financial service providers such as PayPal, Visa, and Mastercard need to stop payments to the manipulation industry. Advertisers need to put sanctions on influencers who use social media manipulation to defraud advertisers, and on social media companies that allow it.

Implications for NATO

Developments in Ukraine, with its advanced antagonistic botfarms, have demonstrated the skill and determination of antagonists seeking to undermine and exploit social media conversations. Disclosures by social media companies underscore the intensity and determination of foreign states and other antagonists to undermine the interests of the Alliance.

Social media manipulation continues to be a challenge for NATO; it is a potent tool for malicious actors seeking to undermine 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 exploit social media for malign 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.

The recent assessment published by the European Union underscores the fact that the ability of the social media companies to protect users in all countries and languages of the Union isn't evenly distributed. In fact, there is good reason to assume that some languages and regions of the Alliance are virtually unprotected by human moderators. NATO must demand full disclosure from social media companies about their intention and ability to ensure that all of the Alliance is protected against hostile foreign manipulation.[42]

ENDNOTES

1. cf. European Commission, 'Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions', 3 December 2020.
2. Facebook, 'Coordinated Inauthentic Behavior Archives'. [Accessed 30 October 2020]'
3. Google, 'Threat Analysis Group (TAG)'. [Accessed 30 October 2020].
4. Twitter, 'Twitter Safety'. [Accessed 30 October 2020].
5. European Commission, 'Assessment of the Code of Practice on Disinformation', 10 September 2020.
6. European Commission, 'European Democracy Action Plan: Making EU Democracies Stronger', [Accessed 3 December 2020].
7. SSU, 'СБУ блокувала діяльність потужної «ботоферми», яку координували куратори з Росії'. [Accessed 30 October 2020].
8. SSU, 'У Бердянську СБУ блокувала роботу ботоферми, керованої з РФ'. [Accessed 30 October 2020].
9. SSU, 'СБУ викрила ботоферму, через яку поширювали фейки про COVID-19 та заклики до повалення конституційного ладу (відео)'. [Accessed 30 October 2020].
10. Graphika, 'IRA Again: Unlucky Thirteen'. [Accessed 30 October 2020].
11. Stanford Internet Observatory, 'Hacked and Hoaxed: Tactics of an Iran-Linked Operation to Influence Black Lives Matter Narratives on Twitter'. [Accessed 30 October 2020].
12. Facebook, 'Removing Coordinated Inauthentic Behavior', 8 October 2020.
13. Stanford Internet Observatory, 'Analysis of an October 2020 Facebook Takedown Linked to U.S. Political Consultancy Rally Forge'. [Accessed 30 October 2020].
14. Stanford Internet Observatory, 'Analysis of an October 2020 Facebook Takedown Linked to the Islamic Movement in Nigeria'. [Accessed 30 October 2020].
15. NATO StratCom CoE, 'Robotrolling 2020/3'
16. See, for example: Ferrara, Emilio, Herbert Chang, Emily Chen, Goran Muric, and Jaimin Patel. 'Characterizing Social Media Manipulation in the 2020 U.S. Presidential Election', 19 October 2020.
17. Sebastian Bay and Rolf Fredheim, How Social Media Companies Are Failing to Combat Inauthentic Behaviour Online, 2019.
18. Statista, 'Most Used Social Media 2020'. [Accessed 21 November 2020].
19. NATO StratCom CoE, The Black Market for Social Media Manipulation, December 2018.
20. Sebastian Bay and Rolf Fredheim, 'How Social Media Companies are Failing to Combat Inauthentic Behaviour Online', 2019.
21. 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).
22. Facebook, Community Standards: 'Inauthentic Behaviour'; Facebook, 'An Update On Information Operations On Facebook', 6 September 2017.
23. See, for example: Facebook, 'White Hat Programme'. [Accessed 21 November 2020].
24. At present we are unable to track bot survival rates on YouTube and TikTok.
25. TikTok, 'Transparency Center'. [Accessed 1 November 2020].

26. Twitter, 'Twitter Transparency Center'. [Accessed 1 November 2020].
27. Facebook, 'Facebook Transparency'. [Accessed 1 November 2020].
28. Google, 'Google Transparency Report'. [Accessed 1 November 2020].
29. Facebook, 'Facebook Transparency Report | Community Standards', [Accessed 1 November 2020].
30. Twitter, 'Platform Manipulation—Twitter Transparency Center', [Accessed 1 November 2020].
31. Google, 'Threat Analysis Group (TAG)', blog.google. [Accessed 1 November 2020].
32. Twitter, 'Twitter Safety'. [Accessed 30 October 2020].
33. Facebook, 'Coordinated Inauthentic Behavior Archives'. [Accessed 1 November 2020].
34. Tiktok, 'Newsroom | Safety'. [Accessed 1 November 2020].
35. Facebook, 'Community Standards Enforcement Report, November 2020'. [Accessed 19 November 2020].
36. Facebook, 'Facebook Transparency Report | Community Standards'. [Accessed 22 November 2020]
37. Facebook, 'New Facebook and Instagram Research Initiative to Look at US 2020 Presidential Election'. [Accessed August 2020].
38. YouTube, 'YouTube Security & Election Policies—How YouTube Works'. [Accessed 10 November 2020].
39. Facebook, 'Helping to Protect the 2020 US Elections', 21 October 2019. [Accessed 10 November 2020].
40. Twitter, 'Additional Steps We're Taking Ahead of the 2020 US Election'. [Accessed 10 November 2020].
41. European Commission, 'Assessment of the Code of Practice on Disinformation', 10 September 2020.
42. Ibid.

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