Twitter, YouTube ignore takedown requests by the Ukrainian Government

In the wake of Russia's full-scale invasion of Ukraine, Big Tech companies announced they would intensify their efforts to cooperate with the Ukrainian Government to mitigate Russia's information warfare. While the substance and effect of specific measures are unknown to the public, the overall efforts usually include the creation of so-called escalation channels, by which the platforms prioritise the moderation of content flagged by designated partners, such as the Ukrainian Centre for Strategic Communications and Information Security (UCSCIS), an official Ukrainian agency established under the Ministry of Culture.[1] Since the outbreak of the war, the Centre has regularly sent Big Tech companies datasets with content and profiles that, in the Centre's opinion, violate platform terms of service and pose acute threats to the security of individuals or the public through their dissemination of Russian war propaganda, hate speech, or inciting language.

The analysis presented in this report offers unique insights into the effectiveness of measures taken by Big Tech in response to requests by the Ukrainian Government. Our researchers analysed a selection of the datasets of content and profiles flagged by the UCSCIS in order to assess how effectively Big Tech companies were responding to individual flags (see Annex 1).[2] In particular, our researchers analysed the continued availability of:

1. accounts propagating Kremlin war propaganda and hate speech;
2. account impersonations on Instagram and Facebook;
3. Kremlin war propaganda on LinkedIn;
4. hate speech on Facebook, YouTube, and Twitter; and
5. ads that constitute war propaganda on Meta products.

Analysis of the available data suggests that Meta has responded fairly effectively to content flagged by the Ukrainian Government, though the majority of accounts disseminating such content have been permitted to remain on the platform. By contrast, the response rate to flagged content by YouTube, Twitter, and LinkedIn is significantly lower. In addition, staffers working for the Ukrainian Government report that at times it takes up to several weeks for platforms to respond to individual flags, and often there is no response whatsoever. More than three months into the war, officials similarly note that Big Tech companies are still not engaging in structured dialogue with the Ukrainian Government or civil society, and all relevant content and integrity policy decisions continue to be taken by US-based teams, which lack insight into local context in Ukraine and are therefore insufficiently responsive to emerging threats. As a result of these shortcomings,

[1] The Centre was established under the Ministry of Culture and Information Policy of Ukraine as one of the mechanisms for countering disinformation through joint efforts of the state and civil society. The Centre is focused on mitigating external threats, in particular information attacks from the Russian Federation.

[2] We do not claim that the Stratcom Centre's requests caused the removal, nor that all flagged content violated the terms of service of the respective platforms. The analysis was conducted 14-23 June 2022 from a Ukrainian IP address.

this report offers a list of eight recommended actions Big Tech companies could take to improve their effectiveness in mitigating the threat and impact of Russia's information warfare going forward.

1. Accounts propagating Kremlin war propaganda and hate speech

We analysed the availability status of accounts that posted content flagged as Kremlin war propaganda and hate speech in an attempt to evaluate the degree to which platforms took action against the violators. The graph below shows that the majority of the accounts responsible for reported content remain active as of the date of this report's publication. On a proportional basis, platforms removed more accounts responsible for distributing Kremlin war propaganda than accounts responsible for propagating hate speech.

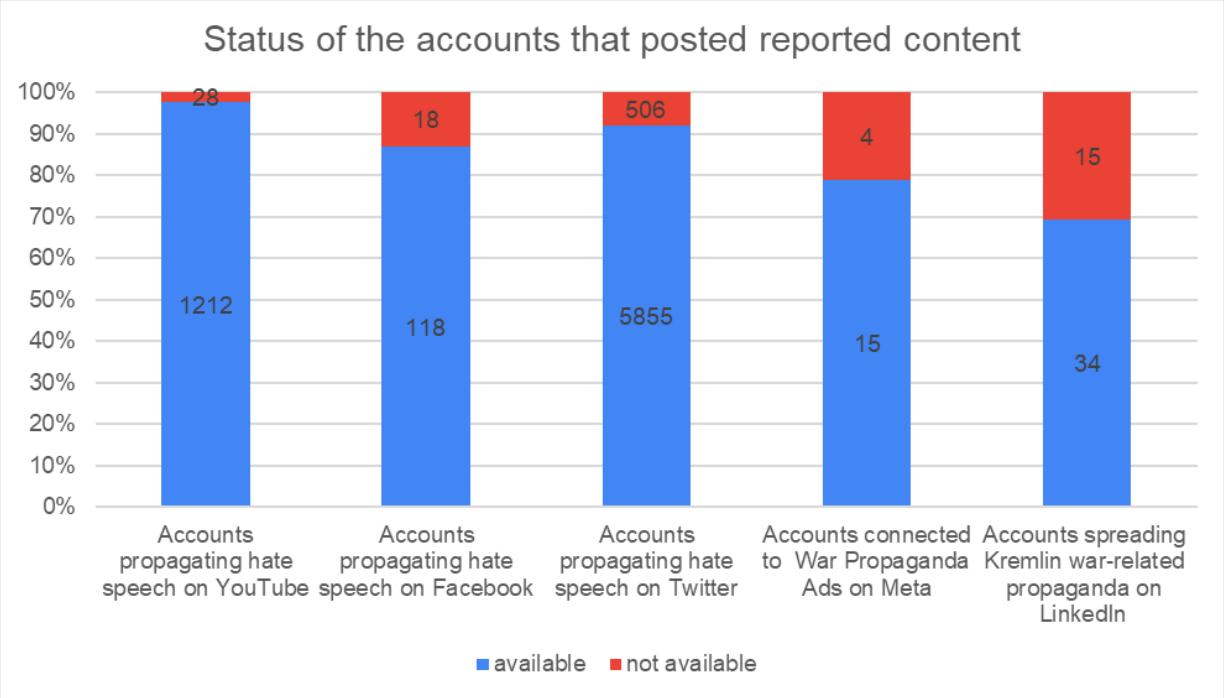


Figure 1: Availability status of accounts that posted reported content.

2. Account impersonations on Facebook and Instagram

We next assessed the availability of accounts flagged for impersonation on Facebook and Instagram. The impersonation samples consist of accounts purporting to be selected Ukrainian government officials and agencies, such as President Zelensky, Mayor of Kyiv Klytchko, Minister of Foreign Affairs Kuleba, the Ministry of Defence, or the Security Service. Imposter accounts can pose a direct threat to the audience, especially in a country with an active conflict, as users may mistake content from an imposter account as official, trustworthy information.

Our analysis found Facebook removed significantly more accounts than Instagram with a margin of 40% of reported accounts. The raw numbers indicate that account impersonations occur more often on Instagram, with 89% more Instagram imposter accounts detected.

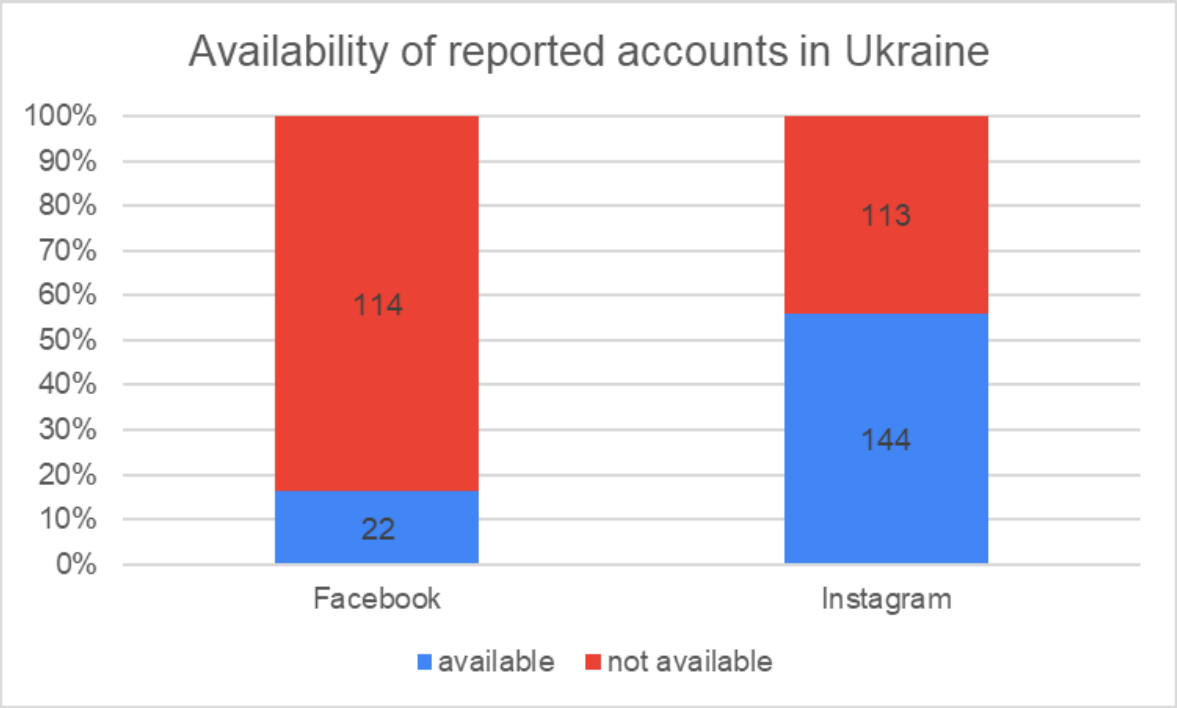
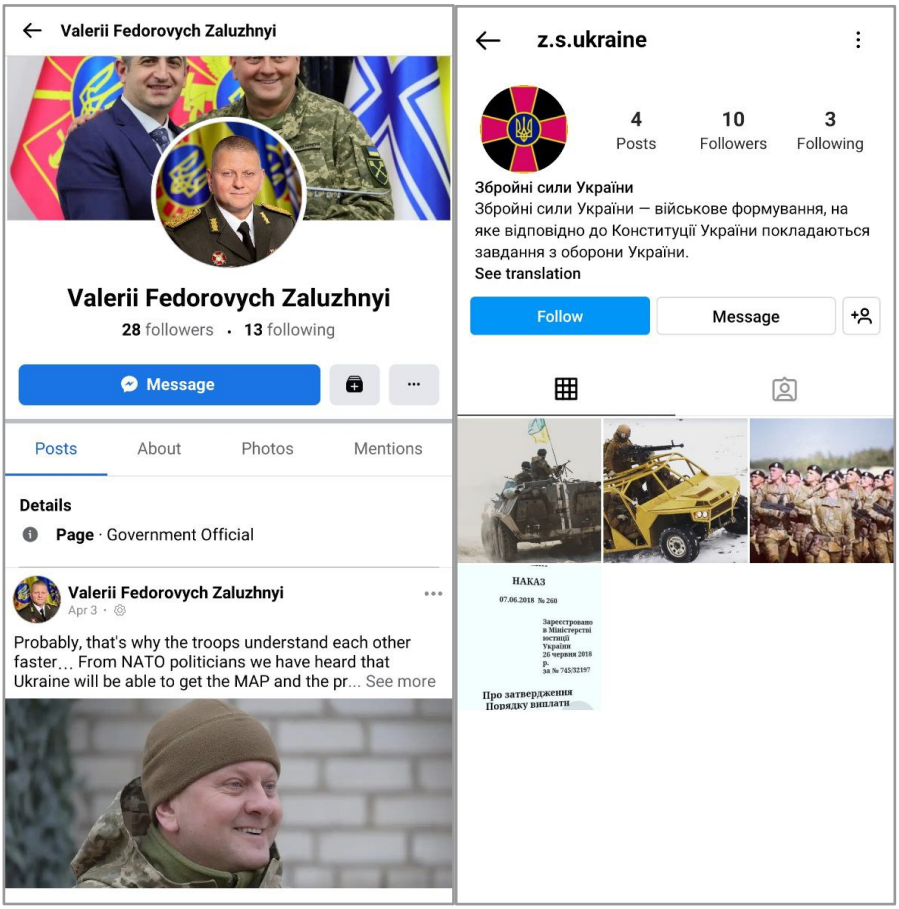


Figure 2: Availability status of reported accounts in Ukraine.

Further analysis of the subset of available accounts revealed that some of the accounts do not conduct any activity and contain only one indicator of possible impersonation, such as the name of the page, while others appear to be "fan pages" or individuals with the same name as the public figure. Other accounts, however, clearly impersonate officials. Our primary concern with these ambiguous accounts is that even if they are dormant or seemingly harmless, they can be mobilised at any time.

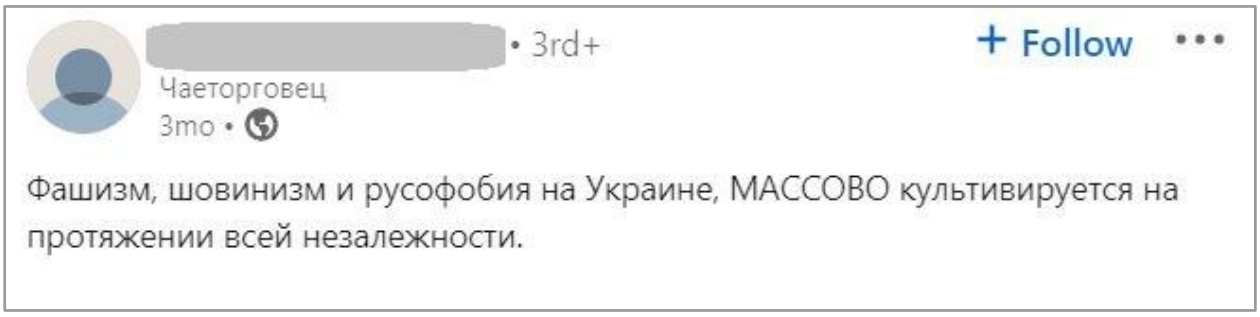
Notably, we also observed that some of the Facebook impersonation accounts are not available in Ukraine, but are available in the EU. This variance raises questions about the consistency of impersonation policy application and the justification behind any such rule (if an account impersonates someone, how/why does location matter).



Example 1. On the left: a Facebook account impersonating Valerii Zaluzhny, the Commander-in-Chief of the Armed Forces of Ukraine (the page is restricted in Ukraine but available in the EU). On the right: an Instagram account impersonating the Armed Forces of Ukraine.

3. Kremlin war propaganda on LinkedIn

Our analysis revealed that 34 of the 65 LinkedIn posts flagged for review by the Ukrainian Government are still accessible on the platform. Many of these posts spread false or misleading content in an attempt to justify Russia's military aggression. One such post that remains on the platform – despite having been flagged to LinkedIn by the Ukrainian Government – claims that the country's independence has allegedly led to a massive rise of fascism, chauvinism, and Russophobia (Example 2, below). Another post promotes a disinformation narrative forged by the Russian Ministry of Defence, according to which secret US biolabs in Ukraine were used to cultivate dangerous pathogens (Example 3, below). Notably, the post at issue contains a link to a YouTube video, which has been already removed.



Example 2. A LinkedIn post. Translation: *"Fascism, chauvinism and Russophobia in Ukraine, massively cultivated throughout the independence."*

Example 3. A LinkedIn post that leads to the deleted YouTube video *"U.S. Biolabs in Ukraine Studied Ways to Transmit Pathogens with the Help of Birds - Russia 24"*.

4. Hate speech on Facebook, YouTube, and Twitter

The samples of hate speech reviewed in our analysis focused predominantly on derogatory terms referring to Ukrainians (such as “ukronazis” or “kh0khols”, see Annex 2). Notably, the words expressing hate in one context may be used as satire in another, so automated processes do not always result in accurately flagged comments (see Annex 3). Nonetheless, the graph below provides an indication of the degree to which platforms have effectively (or ineffectively) responded to content flagged by the Ukrainian Government.

Per our analysis, Facebook removed the majority of the reported content (taking down all reported posts and 83% of reported comments). By contrast, YouTube and Twitter left approximately two-thirds of the reported content up on their platforms. Twitter removed less than one-third of the reported content, despite the fact that the overall volume of problematic content was by far the largest on its platform (there were more instances of reported content on Twitter than on YouTube and Facebook combined).

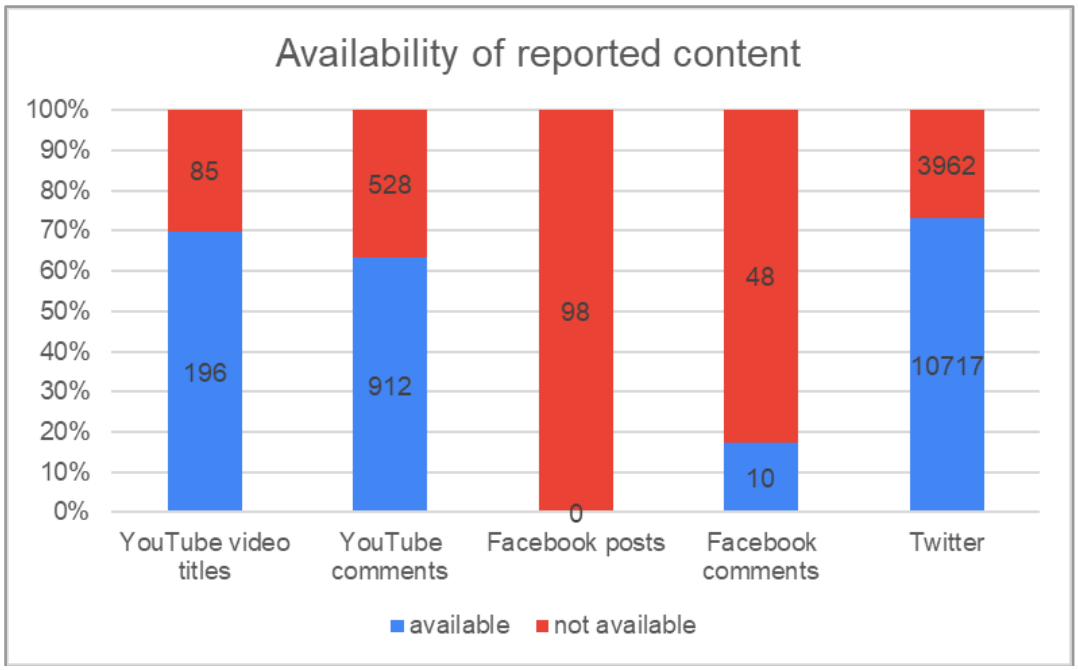
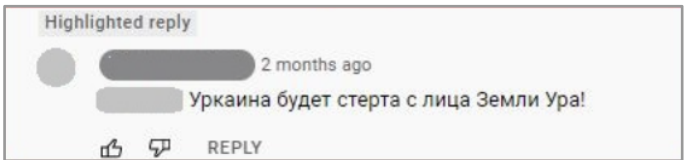
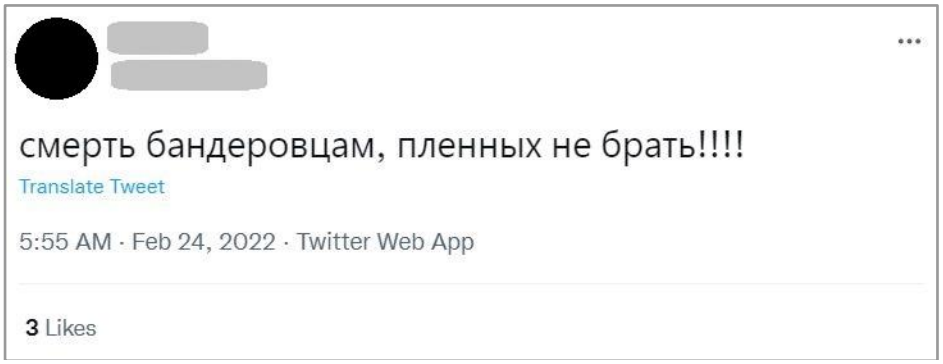


Figure 3: Availability status of reported content on social media platforms.



Example 4. A hate speech comment on YouTube. Translation: *“Urkaina [sic] will be wiped off the face of the Earth Hurray!”*

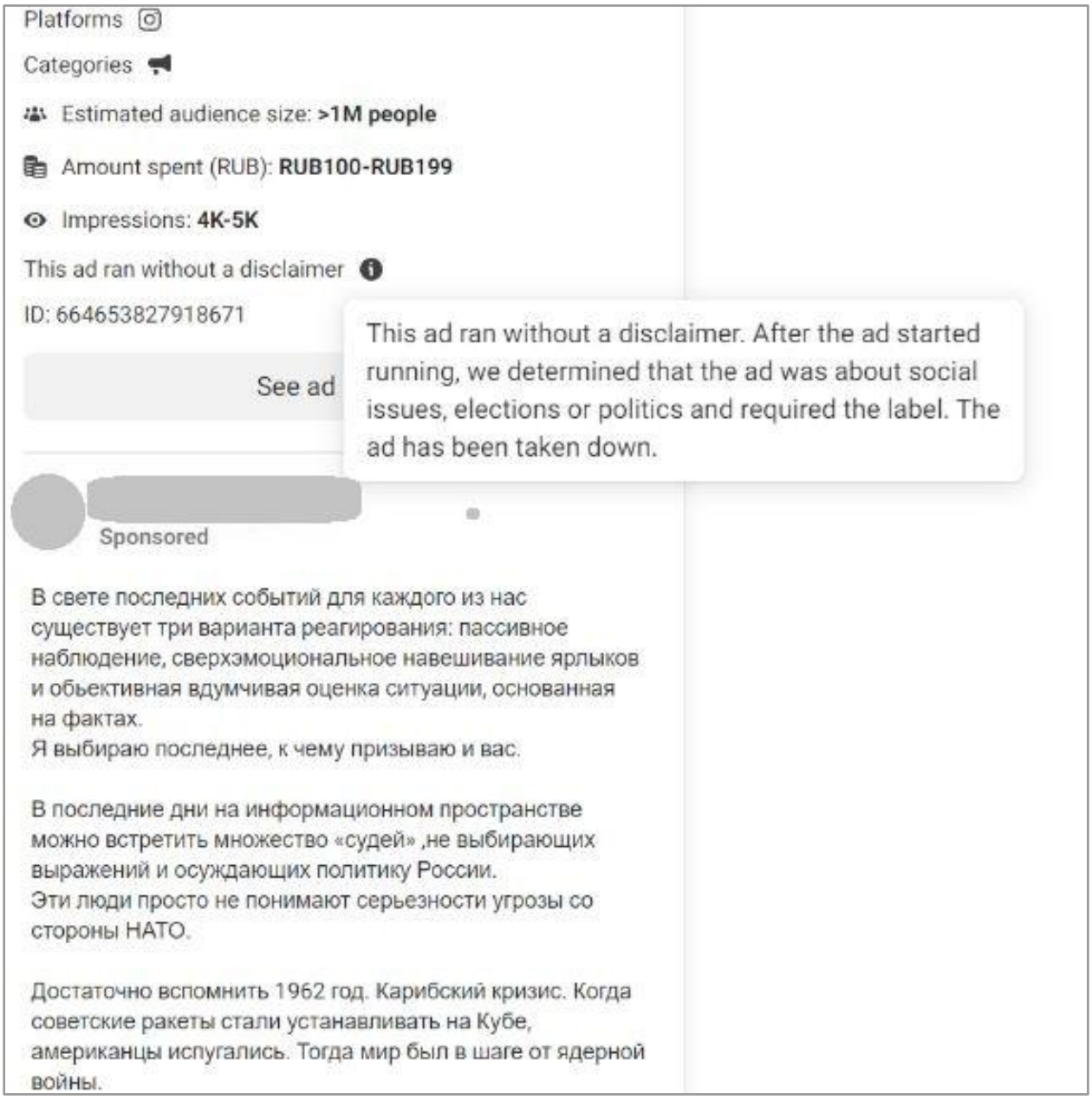


Example 5. A hate speech tweet on Twitter. Translation: *“death to Bandera supporters, do not take prisoners!!!!”*

Several of the comments reported to YouTube appeared to be duplicates. Annex 5 includes an example of such a comment that was left at least by two separate users a total of sixteen times, under six different videos. The comment is a conspiracy claiming that the war is beneficial to the military corporations allegedly controlled by Jews that are to receive money from the US Government. Although in that specific example the derogatory term “ukrofascists” is likely not intended to offend, the comment nonetheless represents a violation of the platform's policies on spam and possible coordinated inauthentic behaviour.

5. Ads that constitute war propaganda on Meta products

Our preliminary analysis determined that all of the ads that were flagged and submitted to Meta were removed by the platform. Upon further review, however, one out of the nineteen reported ads is now visible in the ads library again. The ad is an Instagram post of a Russian female blogger and psychologist who interprets the events surrounding the Russian invasion as NATO’s fault and Russia’s need for self-defence. According to Meta, the ad initially ran without a disclaimer, but was later taken down after Meta determined that the ad’s content was about the war. Although the post is no longer sponsored, it can still be found on Instagram and Facebook (despite the nature of its content).



Example 6. An ad available in the Meta's library justifying the war in Ukraine. Translation of excerpt: "*we will not be able to repel the attack if NATO missiles fly from Ukraine. 5 minutes, and we will be crushed. The promotion of military infrastructure to the East is nothing but preparation for war. [...] Therefore, these are forced measures of influence. This is not a war with Ukraine! We are fighting for our freedom. Of course, war is terrible. God forbid there will be huge human losses. But to say today, 'We are for peace' means to say, 'We are for surrender'*".

6. Recommendations for Big Tech companies

In light of the shortcomings identified in this report, Big Tech companies could take the following actions to improve the effectiveness of their efforts to mitigate the threat and impact of Russia's information warfare as it relates to the war in Ukraine:

1. **Disrupt** Russian hybrid attacks, false flag operations, or coordinated trolling attacks within one hour of receiving notifications by competent Ukrainian authorities or civil society organisations. Accounts that share Kremlin propaganda content (other than for journalistic purposes), threaten Ukrainians, or justify the war on false pretences should be disabled.
2. **Enforce** mitigation measures – such as post removal, account suspension, and algorithmic deprioritisation – in a proactive manner across all content that is identical or very similar to the type/form/origin of content which has been mitigated in the past.
3. **Report** suspected or proven breaches of Ukrainian law, human rights violations, or coordinated disinformation attacks to Ukrainian authorities in real time.
4. **Preserve** content and accounts removed in relation to the war, including any evidence of war crimes and Kremlin-backed information operations, for later use by appropriate Ukrainian authorities.
5. **Protect** Ukrainian users, journalists, politicians, and civil society by fast-tracking notifications from accredited accounts; verifying new accounts or pages containing the names of Ukrainian politicians or institutions before allowing them to go live; and monitoring the accounts of Russian soldiers in Ukraine as well as dormant pages and accounts that were likely created as part of Russia's information operation.
6. **Establish** an early warning system to alert particular groups and individuals exposed to online attacks. This should include a streamlined process for Ukrainian Government and civil society, as well as international partner organisations, to flag offending content or nascent channels.
7. **Secure** the flow of reliable information in Ukraine by calibrating newsfeed algorithms and recommender systems to prioritise engagement signals in favour of verified Ukrainian sources. Verified Ukrainian sources should be exempt from automated bans and suspensions (triggered by malicious user-flagging) and users in areas of active conflict should be directed to authoritative information provided by the Ukrainian Government.
8. **Consult** Ukrainian Government and civil society, as well as their international partner organisations, in a structured format regarding the formulation and execution of policies related to the war (and share attribution and enforcement standards with them). Transparency and access to aggregated data on the views of, and engagement with, high-reach public accounts should be expanded for independent researchers.

ANNEXES

Annex 1. Datasets and availability checks

Datasets provided by the Centre for Strategic Communications and Information Security of Ukraine, and the results of the availability check.

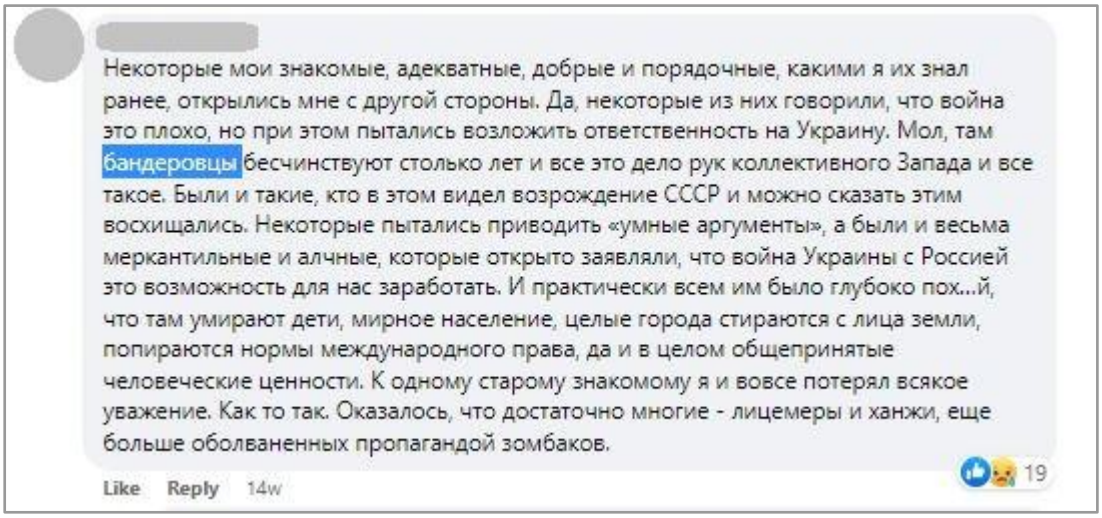
Annex 2. List of keywords used to for search for hate speech content

Keywords	Explanation
укрофашистов, укровшисты, укрофашисты, укровермахт, укронацистам, укро-нацисты, укронацисты	Derogatory name used to call Ukrainians Nazis and fascists
х0хлам, х0хлы, хохол, хохложоп, хохл*	Derogatory name for Ukrainians
укробляди	Derogatory name used to call Ukrainians bitches
бэндеровцы, бэндеровцам, бандеровцам, бандеровцы, бандерлоги, бандеровская	Derogatory name for Ukrainians stemming from the name of Stepan Bandera (a well-known Ukrainian nationalist, portrayed by Russian propaganda as an enemy)
укропам, укропов, укропы, укропия, укропию	Derogatory name for Ukrainians/Ukraine (although it is rarely used in a normal context)
майданутые, майдауны	Derogatory name for Ukrainians, stemming from "Maidan" (referring to the Maidan Uprising of 2013-2014, when Ukrainians fought for the pro-European trajectory of the country's future)
укробабуины	Derogatory name forUkrainians (the word is a combination of “Ukraine” and “baboon”)
уркаина	Derogatory name for Ukraine
гейраина	Derogatory name for Ukraine (the word is a combination of “Ukraine” and “gay”)

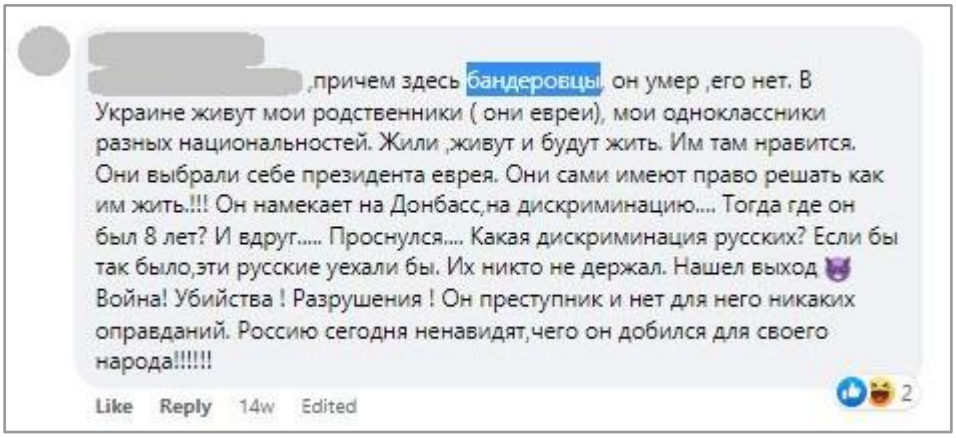
Annex 3. Derogatory terms used without the intent to offend

Example #1 of a derogatory term used without the intent to offend. Translation: “Some of my acquaintances, adequate, kind and decent, as I knew them before, opened up to me from the other side. Yes, some of them said that the war was bad, but at the same time, they tried to blame Ukraine. Like, the Bandera supporters have been rampaging there for so many years, and all this is the work of the collective West and all that. There were those who saw in the war the revival of the USSR and, one might say, admired it. Some tried to give ‘smart arguments’, but there were also very mercantile and greedy ones who openly stated that the war between Ukraine and Russia is an opportunity for us to earn money. And almost all of them did not give a f**k that children, civilians die there, entire cities are wiped off the face of the earth, the norms of international law

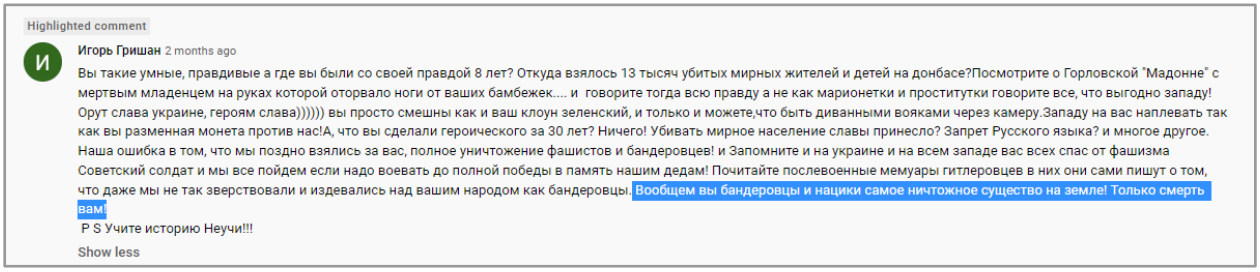
are violated, and, in general, the generally accepted human values. I lost all respect for one old acquaintance. Something like this. It turned out that quite a few were hypocrites and hypocrites, and even more zombies fooled by propaganda."



Example #2 of a derogatory term used without the intent to offend. Translation: "what Bandera supporters have to do with it, he died, he is gone. My relatives live in Ukraine (they are Jews), my classmates are of different nationalities. Lived, live and will live. They like it there. They chose a Jewish president. They have the right to decide how they live.!!! He hints at the Donbas, at discrimination.... Then where was he for 8 years? And suddenly ..... I woke up .... What kind of discrimination against Russians? If that were the case, these Russians would have left. Nobody was holding them. Found a way out 😈 War! Murder! Destruction! He is a criminal, and there is no excuse for him. Today they hate Russia; what did he achieve for his people !!!!!!"



Annex 4. Examples of hate speech comments


Highlighted comment

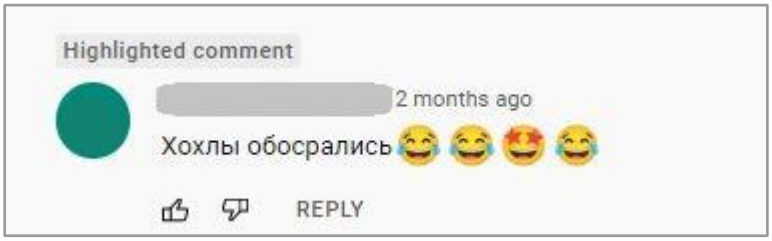
И Игорь Гришан 2 months ago

Вы такие умные, правдивые а где вы были со своей правдой 8 лет? Откуда взялось 13 тысяч убитых мирных жителей и детей на донбасе?Посмотрите о Горловской "Мадонне" с мертвым младенцем на руках которой оторвало ноги от ваших бамбежек.... и говорите тогда всю правду а не как марионетки и проститутки говорите все, что выгодно западу! Орут слава украине, героям слава)))))) вы просто смешны как и ваш клоун зеленский, и только и можете,что быть диванными вояками через камеру.Западу на вас наплевать так как вы разменная монета против нас!А, что вы сделали героического за 30 лет? Ничего! Убивать мирное население славы принесло? Запрет Русского языка? и многое другое. Наша ошибка в том, что мы поздно взялись за вас, полное уничтожение фашистов и бандеровцев! и Запомните и на украине и на всем западе вас всех спас от фашизма Советский солдат и мы все пойдем если надо воевать до полной победы в память нашим дедам! Почитайте послевоенные мемуары гитлеровцев в них они сами пишут о том, что даже мы не так зверствовали и издевались над вашим народом как бандеровцы. Вообщем вы бандеровцы и нацики самое ничтожное существо на земле! Только смерть вам!

P S Учите историю Неучи!!!

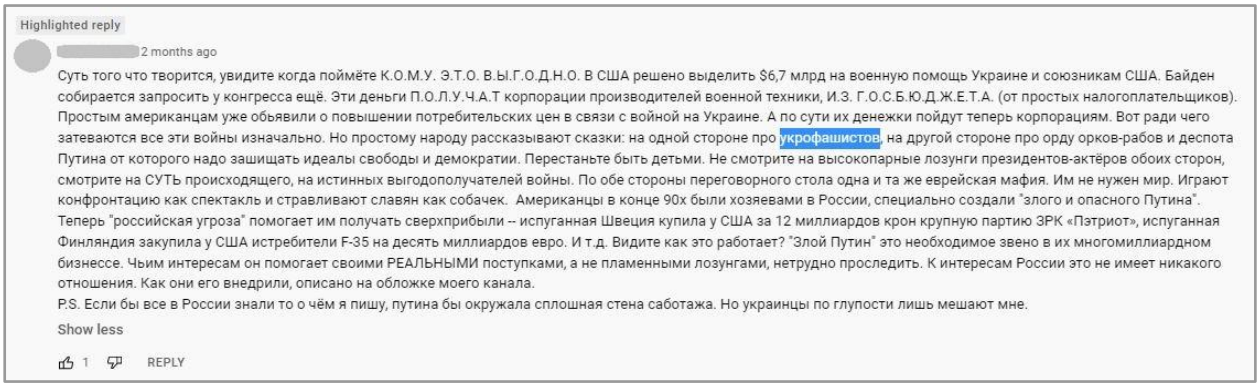
Show less

Example #1 of a hate speech comment on YouTube (still present on the platform). Translation of the highlighted portion: “In general, you Bandera and Nazis are the most insignificant creature on earth! Only death should come to you!”



Example #2 of a hate speech comment on YouTube (still present on the platform). Translation: “Khokhols shit themselves😂😂😂😂”

Annex 5. Example of spam and possible CIB comments


Highlighted reply

2 months ago

Суть того что творится, увидите когда поймёте К.О.М.У. Э.Т.О. В.Ы.Г.О.Д.Н.О. В США решено выделить $6,7 млрд на военную помощь Украине и союзникам США. Байден собирается запросить у конгресса ещё. Эти деньги П.О.Л.У.Ч.А.Т корпорации производителей военной техники, И.З. Г.О.С.Б.Ю.Д.Ж.Е.Т.А. (от простых налогоплательщиков). Простым американцам уже обьявили о повышении потребительских цен в связи с войной на Украине. А по сути их денежки пойдут теперь корпорациям. Вот ради чего затеваются все эти войны изначально. Но простому народу рассказывают сказки: на одной стороне про укрофашистов, на другой стороне про орду орков-рабов и деспота Путина от которого надо зашищать идеалы свободы и демократии. Перестаньте быть детьми. Не смотрите на высокопарные лозунги президентов-актёров обоих сторон, смотрите на СУТЬ происходящего, на истинных выгодополучателей войны. По обе стороны переговорного стола одна и та же еврейская мафия. Им не нужен мир. Играют конфронтацию как спектакль и стравливают славян как собачек. Американцы в конце 90х были хозяевами в России, специально создали "злого и опасного Путина". Теперь "российская угроза" помогает им получать сверхприбыли -- испуганная Швеция купила у США за 12 миллиардов крон крупную партию ЗРК «Пэтриот», испуганная Финляндия закупила у США истребители F-35 на десять миллиардов евро. И т.д. Видите как это работает? "Злой Путин" это необходимое звено в их многомиллиардном бизнессе. Чьим интересам он помогает своими РЕАЛЬНЫМИ поступками, а не пламенными лозунгами, нетрудно проследить. К интересам России это не имеет никакого отношения. Как они его внедрили, описано на обложке моего канала.

P.S. Если бы все в России знали то о чём я пишу, путина бы окружала сплошная стена саботажа. Но украинцы по глупости лишь мешают мне.

Show less

1 REPLY

An example of a comment on YouTube promoting conspiracy and having signs of being amplified inauthentically (the alleged conspiracy relates to the war being beneficial to the military corporations controlled by Jews that will receive money from the US state budget). Translation: “You will see the essence of what is happening when you understand W.H.O B.E.N.E.F.I.T.S. OUT OF IT. In the United States, it was decided to allocate $6.7 billion for military assistance to Ukraine and US allies. Biden is going to ask Congress for more. This money W.I.L.L. R.E.C.E.I.V.E the corporations manufacturing military equipment, F.R.O.M. T.H.E. S.T.A.T.E.B.U.D.G.E.T. (from ordinary taxpayers). An increase in consumer prices in connection

with the war in Ukraine has been already announced to ordinary Americans. And, in fact, their money will now go to corporations. That's why all these wars are planned from the very beginning. But fairy tales are told to the common people: on the one hand, about the Ukrofascists, on the other hand, about a horde of orc-slaves and despot Putin, from whom the ideals of freedom and democracy must be protected. Stop being kids. Do not look at the grandiloquent slogans of the presidents-actors of both sides; look at the ESSENCE of what is happening, at the true beneficiaries of the war. On both sides of the negotiating table is the same Jewish mafia. They don't want peace. They play confrontation-like performances and pit the Slavs like dogs. The Americans in the late 90s were the masters in Russia; they created the ‘evil and dangerous Putin’ on purpose. Now, the ‘Russian threat’ is helping them to make super profits - frightened Sweden bought a large batch of Patriot air defence systems from the USA for 12 billion crowns, and frightened Finland bought F-35 fighter jets from the USA for ten billion euros. Etc. See how it works? ‘Evil Putin’ is a necessary link in their multi-billion dollar business. Whose interests he helps with his REAL actions, and not with fiery slogans, is easy to trace. It has nothing to do with Russia's interests. How they are implemented is described on the cover of my channel. P.S. If everyone in Russia knew what I am writing about, Putin would be surrounded by a solid wall of sabotage. But the Ukrainians foolishly only interfere with me.”