The Bratislava Shooting

Report on the role of online platforms

COUNCIL FOR MEDIA SERVICES

Reset.

Photo by Jeremy Bishop

I. Overview & Key Findings

On 12 October 2022, a radicalised attacker shot and killed two people and injured another outside an LGBTQ+ bar in Bratislava, Slovakia. Prior to the attack, the shooter shared his manifesto – which contained extremist ideology, hate speech, and harmful conspiracy theories – on multiple file-sharing platforms. After the shooting spree, the attacker took to posting on Twitter and 4chan about the event before committing suicide. Local authorities subsequently reclassified the attack from murder to an act of terrorism, and debate over the incident soon spread rapidly on social media with many users posting hateful comments aimed at further injuring the LGBTQ+ community.

In an effort to mitigate future damage and incitements to violence after the attack, the Slovak regulator – the Council for Media Services (CMS) – sought to collaborate with Facebook, YouTube, and Twitter by flagging content and engaging in intensive bilateral communications. Despite the urgency, the platforms's content moderation efforts were very slow, even in clear cases.

Given this troublingly low degree of cooperation, this report analyses the content moderation policies and responsiveness of Twitter, Facebook, Instagram, and YouTube in the context of the shooting in Bratislava. Particular insight was gained for the report by collaborating with the CMS in order to share process details and results, and to push for greater oversight of the analysed platforms. Key findings are as follows:

- **Content Moderation System Failures:** The content moderation systems of the analysed platforms failed to identify extremist, terrorist, and hateful content both before and after the terrorist attack in Bratislava, regardless of language used.
    - *Pre-Attack:* Even though these systems perform best when analysing English-language content, Twitter's content moderation tool failed to detect problematic content published by the terrorist prior to his attack – despite the posts clearly violating both platform Terms of Service (ToS) and local law.
    - *Post-Attack:* Platforms like Facebook, Instagram, and YouTube failed to sufficiently moderate hateful content after the attack and efficiently enforce its content policies. On these platforms, more than 10% of the 300 most-toxic comments detected under posts related to the terrorist attack contained hate speech against the LGBTQ+
- **Slow and Insufficient Platform Responses:** According to the CMS, Facebook's response to the national regulator's inquiries to remove inciting content related to the attack was slow and contradictory. In the first three weeks after the shooting, the CMS flagged 66 relevant posts that it deemed to be in violation of Facebook's community standards. Facebook removed only six of the flagged posts, with an average response

time of eleven days, and failed to provide any reasoning as to why the remaining content was not removed.

- **Repeated ToS Offenders Not Penalised:** The most frequent authors of potentially illegal or harmful content were "repeat offenders" who had recurrently violated platform Terms of Service previously and repeatedly been reported by the CMS prior to the terrorist attack.

- **Insufficient Content Moderation Resources:** The problematic content reported by the CMS was usually sent to a third-party fact-checker hired by Facebook to perform the review. In a stark signalling of market priority, there is only one Facebook-contracted fact-checker for all of Slovakia. The limited resources Facebook dedicates to fact-checking in small markets poses yet another obstacle for the quick and efficient review and removal of potentially harmful content.

- **Outdated Counter-Terrorism Policies:** Platform policies on extremism and terrorism differ from platform to platform, resulting in inconsistent content moderation policy application on posts that can relate to life-and-death situations. Worse, current platform ToS tend to focus on terrorist organisations despite the recent surge in attacks perpetrated by lone actors not formally affiliated with a specific terrorist organisation. While both Meta and Twitter do account for such a possibility in their respective ToS, the scope of existing content moderation policies that address lone actors is far too limited and vague when compared to those that apply to terrorist groups. For example, in Meta's policy on terrorism are the only references to lone actors in the examples, and not in the actual wording of the policy. Together, these failures work to stymie counter-radicalization and counter-terrorism efforts.

- **Inefficient Content Moderation Poses Systemic Risk:** Major mainstream social media platforms have repeatedly failed to remove illegal content and other content in clear violation of terms of service even when it was proactively reported by users, as demonstrated in Italy, Germany, and France. The inability to prevent the spread of terrorist content online is not limited to the instant case, but rather represents a pervasive systematic problem embedded in the functioning of social media platforms, especially when it comes to smaller markets deemed to be politically less significant.

In the sections that follow, this report provides further background regarding the Bratislava terrorist attack and the social media platforms that were utilised by the shooter prior to and after the assault. The role of the CMS as the designated national regulatory authority in Slovakia is then discussed before moving to an evaluation of platform policy and policy implementation in the context of the shooting. Next, an analysis of platform cooperation with the CMS is offered – highlighting disparities in treatment for small-market countries and inadequacies in compliance with the future obligations of the Digital Services Act. The report concludes with a series of lessons learned and recommendations intended to tackle societal risks, including radicalisation and polarisation, and improve content moderation efficacy of digital platforms.

II. Bratislava Terrorist Attack

a. Initial assault

On the evening of 12 October 2022, a shooting attack was carried out in front of an LGBTQ+ bar located at the Zámocká street in the centre of Bratislava, Slovakia. During the attack, two people were killed and another injured. The shooter escaped immediately after the incident and remained at large until the next morning, when the alleged attacker was found dead after commiting suicide.

Shortly after the attack, information about a Twitter account allegedly belonging to the attacker started to spread across social media platforms. The account in question tweeted after the attack about the incident and declared “feeling no regrets” in English while using hashtags such as #bratislava, #hatecrime, and #gaybar (Example A). As midnight approached on the night of the attack, the same account posted a tweet in Slovak suggesting the user would commit suicide (Example B). At the same time that the alleged attacker was tweeting these comments on Twitter, he was actively posting about the incident on 4chan and confirming his identity as the perpetrator of the attack. On both platforms, the individual communicated predominantly in the English language, with only a few posts made in Slovak.

Example A: *The shooter’s tweets shortly after the attack.*

Example B: *The shooter’s final tweet. English translation: “Bye, see you on the other side”.*

Notably, five hours before the shooting began, the alleged attacker also shared via Twitter several URLs to his 65-page-long manifesto which included references to antisemitic conspiracy theories, extremist ideologies, militant accelerationism, and previous terrorist attacks in Buffalo, New York, and Christchurch, New Zealand.

Prior content shared on the alleged attacker's Twitter account suggests that the individual prepared in advance for the attack. In August 2022, for example, he shared a picture of himself in front of the LGBTQ+ bar where he would later carry out the attack (Example C). In turn, just a few hours before the attack in October, he tweeted that he had made his "decision" (Example D). When viewed in tandem with the alleged attacker's 4chan activity and extremist manifesto, these Tweets indicate the attack was premeditated.

Example C: *The shooter posting pictures of himself on Twitter in August 2022 in front of the LGBTQ+ bar where he would carry out his attack in October.*

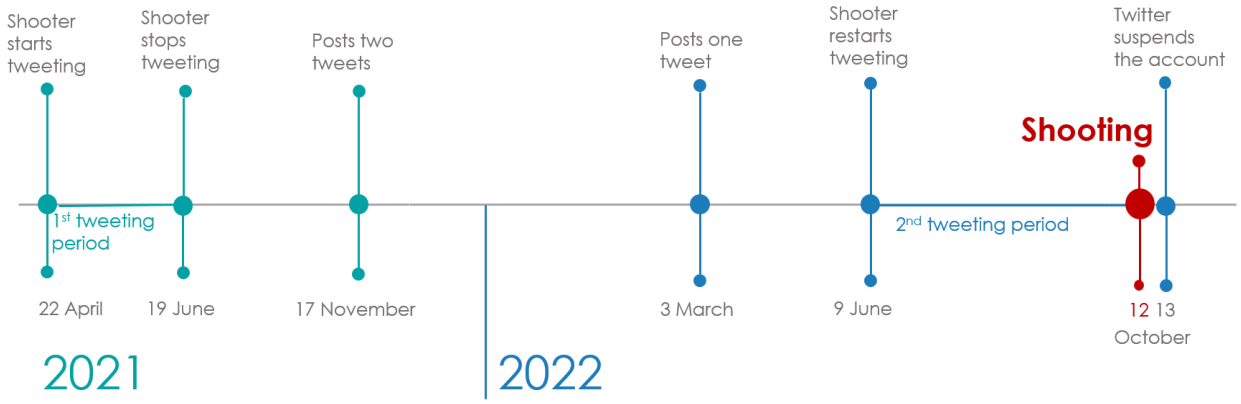
Example D: *Tweets from the shooter on the day of the attack and the day before.*

Subsequent investigation into the events confirmed the identity of the attacker and linked him to the aforementioned Twitter and 4chan accounts. The attacker, a nineteen-year-old man from Bratislava, was the son of a former candidate of the far-right party Vlasť (during the 2020 general elections) and had no known personal connection to the victims of his crime. Five days after the attack, the National Crime Agency and General Prosecutor of Slovakia subsequently reclassified the crime from murder to terrorist attack. Under Slovak law, acts of terrorism are those that are committed with the aim to destabilise the societal, political, and economic fabric of the state while seriously intimidating its population.

b. Platform utilisation by the shooter

1. Twitter

The attacker used two confirmed social media platforms to communicate with the public: Twitter and 4chan (though a third platform, Telegram, was likely also utilised by the attacker to communicate). The attacker established his Twitter account in April 2021 and used it to tweet during two distinct time periods: April 2021 to June 2021, and June 2022 to October 2022. During the year gap in time between these periods, he tweeted only three times: twice on 17 November 2021 and once on 3 March 2022. Data review for this report shows the attacker had very few to no Twitter followers, and the majority of interactions under his tweets occurred after the October attack.



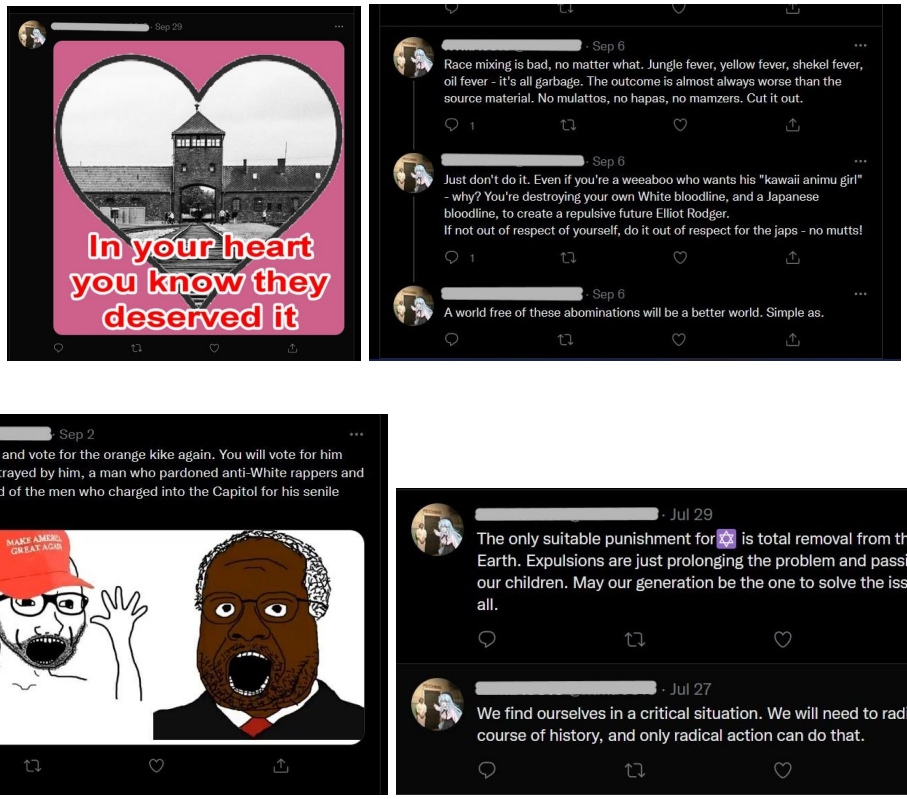
Graph 1: *Timeline of the shooter's tweeting activity.*

From April 2021 to June 2021, he tweeted exclusively in English and used coded language to post hateful content (mostly antisemitic and anti-Black narratives). A few months later, on 17 November 2021, he mentioned "forced vaccination" and suggested causing harm to the people implementing vaccination policies. Similar conspiratorial thinking was later mentioned in his manifesto, which he shared during the day of the shooting.

Example E: *Tweet from the shooter in English mentioning "forced vaccination" and suggesting to cause them harm.*

The attacker's second period of tweeting activity included hateful and extremists memes, tweets using antisemitic and racist slurs, and positive comments about other far-right terrorist attacks. From August 2022, more unambiguously hateful and violent content was posted alongside pictures of himself in front of the LGBTQ+ bar where he would later carry out the attack and the

house of the Slovak prime minister. According to his manifesto and communication on 4chan, these pictures may have been some sort of memorabilia of him planning the attack. While the content posted during this time began as mainly antisemitic and anti-Black hate speech, it turned in September to include more anti-LGBTQ+ content and content glorifying far-right terrorists.

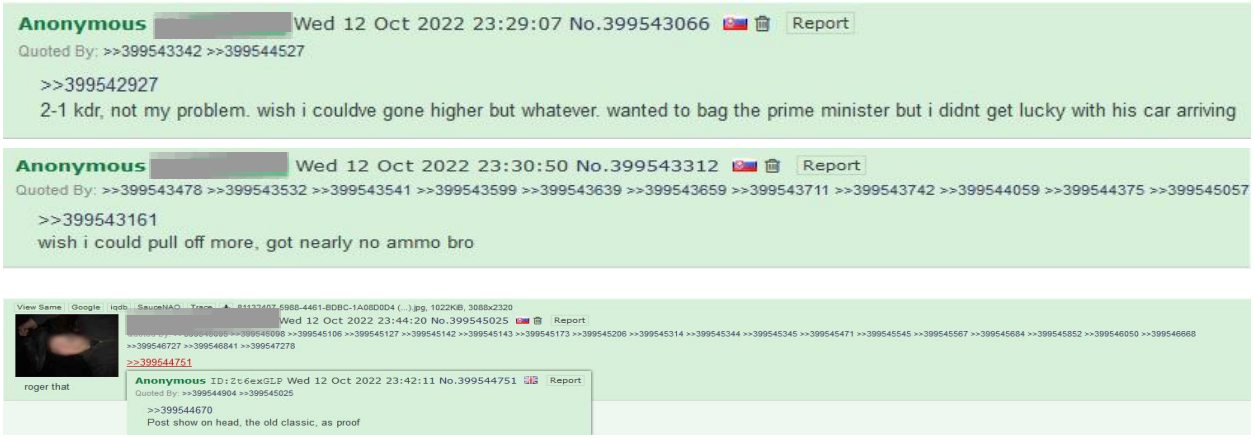


Example F: *Tweets from the shooter containing hate speech, antisemitism and incitement to violence.*

2. 4chan

Unlike prior accelerationists such as the Christchurch and Buffalo shooters, the perpetrator from Bratislava did not livestream his attack. Instead, as noted above, he posted related content and comments on various social media channels after the attack had been carried out. In the wake of his shooting spree, the attacker used 4chan's /pol section to publicly confess to carrying out a terrorist attack. His first post, a comment to a thread warning of a possible terrorist attack in Bratislava, appeared online several hours after the shooting (23:24). After making his first post on 4chan, he began replying to questions posted by other users, casually mentioning his desire to kill the Slovak PM or how a lack of ammunition prevented him from killing more people (Example G). At 23:38, the attacker published his first selfie after being prompted to do so by an anonymous user. Other users subsequently requested pictures of him with a shoe on his head (a potential reference to the alt-right trolling community), or pictures of racist slurs spelled out using leaves from the park in which he was hiding. At 23:52, he posted his last comment

explaining that he had no remorse for carrying out the attack. In total, he posted sixteen times on 4chan in 28 minutes.

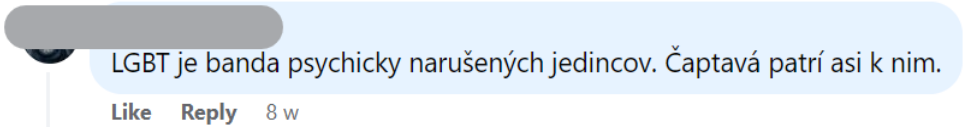


Example G: *A sample of the shooter's conversations with users on 4chan.*

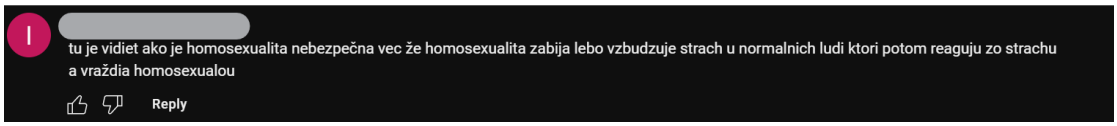
3. Other platforms

According to the attacker's manifesto and Twitter account, content on Telegram was his primary source of inspiration. Given this, the terrorist presumably used the platform to communicate with other Telegram users either before or after the attack, but no information about his Telegram activity has yet been disclosed.

Although the most-commonly-used platforms in the Slovak market like Facebook and YouTube were not used by the terrorist before or after the attack, these platforms subsequently became the primary spaces where online users discussed the crime. In the comments below posts and videos providing information about the attack, researchers for this report detected violent and hateful content against the LGBTQ+ and Jewish communities. Even though both platforms consider the promotion of terrorist acts and hate speech to be a violation of their ToS, they did very little to proactively suppress hostile and illegal content.



Example H: *A hateful comment detected on Facebook after the attack. English translation: "LGBT is a bunch of psychologically ill individuals. Snappy apparently belongs to them" ("Snappy" here is used as word play in reference to the President of Slovakia and her family name).*



Example I: *A hateful comment detected on YouTube after the attack. English translation: "here you can see how homosexuality is dangerous thing and that homosexuality kills because it creates fear by normal people which react from fear and kill homosexuals".*

Example J: *A hateful comment detected on Instagram after the attack. English translation: "fucking faggots".*

III. Role of the Slovak Regulator (CMS)

a. Origin and legal basis

The Council for Media Services (CMS) is the Slovak national regulatory authority responsible for media oversight and enforcement of regulatory frameworks pertinent to broadcasting, retransmission, provision of on-demand audiovisual media services, and digital platforms. The regulator has its origins in the Council for Broadcasting and Retransmission which was reformed in 2022 by the adoption of a new media law, effectively creating a new regulatory body - the CMS. The mission of the Council is to enforce the public interest in the exercise of the right to information, freedom of expression, and the rights of access to cultural values and education.

Under Act no. 264/2022 of the Slovak Republic, the CMS is entrusted with the responsibility and legal competency to prevent the dissemination of illegal content online via systematic oversight over digital platforms. This entails cooperation with the platforms in the effective, proportionate, and non-discriminatory application of their community rules, norms, and standards.

Per article 110 of Act no. 264/2022, the CMS is also tasked with assessing platform content moderation practices and cooperating with online service providers in efforts to enforce their respective community standards or ToS. In this regard, article 151 of Act no. 264/2022 stipulates what constitutes illegal content online (with the applicable definition stemming from the Slovak penal code). Under Slovak law the publication of the following types of content is illegal:

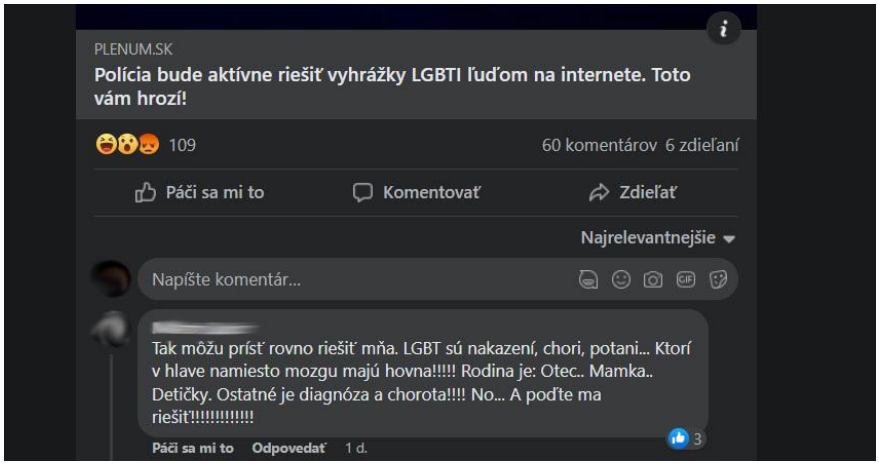
- Content featuring child pornography or extremism.
- Content inciting to violence or featuring acts of terrorism.
- Content approving or praising acts of terrorism.
- Content denying or approving the Holocaust, crimes of political regimes, crimes against humanity, defamation of a nation, race and belief, or incitement to national, racial and ethnic hatred.

b. Interactions with platforms after the Bratislava attack

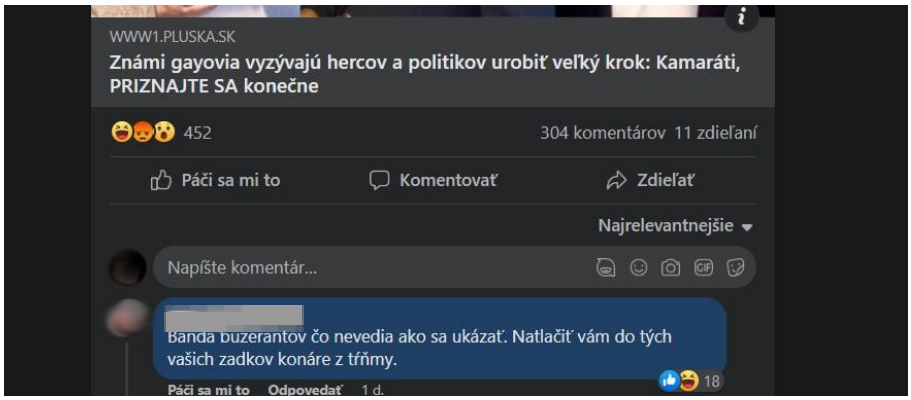
Given the nature of the October 2022 attack, the CMS was immediately and primarily concerned with the dissemination of extremist content online, content approving the acts of terrorism or striving to further incite national, racial, or ethnic hatred.

Immediately after the shooting, the CMS began to use external tools to monitor the online environment – namely Facebook, Twitter, Telegram, YouTube, and specific cloud and storage hosting services. To find potentially harmful or problematic content, random samples were

reviewed based on internal methodologies. In turn, the Council predominantly focused on monitoring Meta's social media platform, Facebook, considering its outsized importance within the Slovak market. During the first three weeks of monitoring, the CMS identified and reported 66 posts and comments related to the attack in Bratislava that it deemed to be in violation of Facebook's community standards. The total number of identified and reported posts by the CMS by 21 October, including other topics such as the war in Ukraine, was 109.



Example K: *A hateful comment detected on Facebook after the attack. English translation: "LGBT people are infected and sick rats… that have shit in their heads instead of brains!!! Family is: Father…Mother..Children. Everything else is a diagnosis and disease!!! So… come at me!!!!"*



Example L: *A hateful comment detected on Facebook after the attack. English translation: "a bunch of faggots who have no other means how to get attention. You should have branches with thorns shoved up your asses."*

The CMS also found ten publicly available online repositories that contained the terrorist's manifesto. In all cases, the CMS contacted the repositories' hosting providers and informed them of the existence of terrorist content, at which time most providers took down the websites containing such content. After locating public versions of the manifesto online multiple times, the Slovak National Crime Agency – in coordination with the CMS – notified the Global Internet Forum to Counter Terrorism (GIFCT) of the file's existence. This information was then hashed

and distributed among GIFCT members. Hashing terrorist content and sharing the hash (a unique digital identifier, like a fingerprint) enables platforms participating in the GIFCT hash-sharing database to quickly identify and remove such content.

As detailed in Section V, below, the CMS also used an escalation channel to communicate with platforms and send inquiries to remove harmful or problematic content.

IV. Evaluation of Platform Policy and Policy Implementation

Although the Bratislava shooting was the first designated deadly terrorist attack in Slovak history, the world writ large has recently experienced a surge in far-right terrorist attacks and live-streamed attacks inspired or fueled by online communities and content. Social media platforms like Facebook have responded to this expansion in terrorist and extremist activity by implementing new policies and content moderation systems designed to prevent the spread of violent content and incitement to violent actions. These policies and their implementation are evaluated below in the context of the terrorist attack in Bratislava. The analysis indicates that problematic and harmful content suggesting a possible attack might occur was not identified in time, and existing platform content moderation and risk mitigation efforts contain gaps in both policy and enforcement.

a. Platform policy

Major social media platforms like Facebook, Instagram, Twitter, and YouTube all have policies in place prohibiting terrorist and extremist content on their platforms, as well as the praise and promotion of such organisations or activities. These policies are regularly updated to reflect new insights and findings.

1. Policy variance

Despite working to address the same systematic problem, policies on terrorism and extremism differ from platform to platform, with differences in definitions, third-party sources, and consequences.

For example, Meta's platforms Facebook and Instagram use a three-tier system which distinguishes between (1) terrorist events and groups that target civilians, (2) entities that engage in violence against state or military actors, and (3) entities that repeatedly engage in violations of Meta's Hate Speech policy or intent to commit a violent attack. Meta allows discussion about terrorist organisations or events only when the discussions are neutral or condemn such events. If a "user's intention is ambiguous or unclear" then the default is to remove the content. Praising or supporting any violent groups or events is, in turn, specifically prohibited by Meta. Tier 1 of the system includes entities that are involved in offline harms, such as terrorists and hate/criminal organisations – as well as their leaders, founders, and prominent members. Entities from the U.S. Government's prescribed list of terrorist organisations are specifically included. When it comes to individual attackers not affiliated with any terrorist organisation, however, the wording of Tier 1 is ambiguous. Although provided examples include instances of ideologically aligned but unaffiliated terrorists (e.g. the attacker from Christchurch,

New Zealand), the policy's wording itself should be clarified to avoid confusion as to whether individuals not specifically aligned with an organisation are indeed covered by the policy.

In turn, Twitter's policy on violent groups is based on promotion or affiliation with illicit activities, and includes terrorist organisations as well as violent extremist groups or individuals. Twitter states that it conducts its own assessment of accounts which might promote terrorism or violent extremism, but no further explanation is provided as to what such a process might entail other than the fact that the assessments are informed by national and international lists of prescribed terrorist and extremist organisations. Although the policy lists the criteria of what constitutes a "violent extremist group" or "other violent organization", no such criteria is provided for individuals "who affiliate with and promote" the illicit activities of the two. In addition to engaging in or promoting violence and targeting civilians, Twitter also considers engaging, supporting, recruiting, or providing services and using insignias of violent organisations to be in violation of this policy (an exception exists for when a group has reformed or the content is used for educational purposes). Despite this, Twitter's policy ultimately lacks clarity regarding how individual attackers will be classified or perceived. The quick action of the platform in the wake of the shooting in Bratislava suggests that the policy does apply to individuals who are unaffiliated with a violent or terrorist organisation, yet Twitter should reflect this reality by updating their public policy (which, as of the timing of this report's publication, was dated October 2020). The internal restructuring of the company following Musk's acquisition may have negatively affected the resources for issues pertaining to safety and security. With even more limited capacities of the Trust and Safety team it will be challenging for the platform to react appropriately in future crisis situations.

YouTube's policy prohibits praise, promotion, or aid to violent criminal organisations. Content is considered to be in violation of the policy if it is produced by criminal or terrorist organisations, praises terrorist or criminal figures, justifies violent acts by violent criminal or terrorist organisations, or depicts symbols of violent criminal organisations. In such situations, YouTube will remove the content and issue a strike against the channel. YouTube's wording of the violent extremist or criminal organisations mentions individual attackers only in the context of praising or memorialising prominent individuals in order to encourage others to carry out acts of violence.

Unlike the above policies from the major digital platforms, alternative online platforms relevant to the Bratislava terrorist attack, such as Telegram and 4chan, have virtually no public policies in place on this critical issue. While Telegram's rules do forbid the promotion of violence on publicly viewable channels, the implication or inference from this rule is that anything goes in closed groups or channels. In the platform's FAQ, Telegram states that it blocks terrorists and acknowledges that competent EU authorities can send a command to remove terrorist content online. 4chan has so-called Global rules and specific rules for each imageboard, yet none of these include any mention of violent, terrorist, or extremist content. The platform rules do, however, state that users cannot upload anything that violates local or U.S. law.

When viewed in the context of the Bratislava shooting, it is apparent that existing platform policies on terrorism and extremism contain serious gaps. Considering that the October attack was carried out by an individual who was not specifically affiliated with any terrorist or violent organisation, the current wording of platform policies against terrorist content suggests that the rules would not – or at least might not –apply to the case. As addressed above, the policies of both Meta and Twitter speak to the actions and content of individual attackers in only a limited way, while YouTube focuses almost exclusively on organisations. These types of gaps create obvious obstacles for actors, such as the CMS, working to prevent problematic content from circulating online either before or after an attack occurs. According to the current wording of YouTube's policy, much of the content produced by the shooter would not be removed – although if the same content was created by an extremist, criminal, or terrorist organisation, the conditions would be sufficient to remove it.

b. Platform policy implementation

1. General inadequacies in automation policy implementation

Public scrutiny of the content moderation policies of large digital platforms has increased significantly in the wake of atrocities like the 2018 Myanmar genocide and the 2019 Christchurch attack. Due in part to this shift, platform efforts to monitor and filter user-generated content to ensure compliance with ToS and community standards are no longer ignored or unreported. In an effort to better address the incredible volume of content shared online, the largest digital platforms have in turn shifted their content moderation efforts towards automation. Major platforms almost always now use algorithmic content moderation as the first step in preventing the spread of illegal and harmful content online. Under this system, after a piece of content is flagged as problematic, it is either sent to a human operator for further review or automatically deleted.

Yet automated content moderation involves numerous pitfalls, ranging from pre-existing bias embedded in data labelling to a general lack of human resources devoted to reviewing flagged content. The disparity in the accuracy success rate between text and image recognition is another pervasive problem of automated content moderation. In the Bratislava case, the attacker posted explicitly violent and illegal textual content on Twitter that should have been detected by Twitter's content moderation tool but was not. Notably, the attacker's tweets were almost exclusively in English – a language automated moderation systems are supposed to evaluate best. The feed of the attacker's account also contained clearly extremist content with pictures depicting nazi-symbols alongside racist and antisemitic slurs.

Example M: *Screenshots of explicitly violent and illegal content posted on Twitter by the attacker that should have been detected by the platform's automated content moderation tool but was not.*

The fact that none of this problematic content was detected by Twitter prior to the attack reinforces the type of questions that have been repeatedly raised by experts regarding the extent and degree to which automated moderation systems actually help to enforce platform policies in a timely and efficient manner. Unfortunately, it often appears that the technological limitations of such systems render them ineffective when they are deployed on a massive scale (resulting in too much content being flagged for human review) or to combat fringe users and publishers of illegal content who work to game the system.

The technological shortcomings of automated content moderation are further exacerbated in lower-priority markets such as Central and Eastern Europe, where efforts to remove content in violation of platform policies are usually slow and inefficient – and largely disregard the local language. For example, in a recent review of content moderation procedures in Central and Eastern European countries, Facebook removed only 44% of content reported in the Slovak language, while YouTube and Twitter removed 16% and 0% respectively. For the majority of cases in the study, Twitter responded that the content did not violate its policies, even though the comments flagged for review included hate speech or incitement to violence. The research also demonstrated inconsistencies in local content moderation as posts that included almost identical content were evaluated differently, and reported comments were permitted to stay up online for as long as 235 days prior to removal. Similar or even worse results were found in all Central and Eastern European countries, with particularly dissatisfying results in Hungary.

2. Policy implementation in response to the Bratislava terrorist attack

In the context of the October 2022 terrorist attack in Bratislava, social media platforms did not respond in a timely manner to the requests made by the CMS to remove problematic content. Facebook removed almost none of the harmful content violating its ToS connected to the terrorist attack for which a notification was issued by the CMS. By the end of October, the CMS detected and flagged 66 posts directly related to the attack that violated Facebook's community standards. A qualitative analysis conducted by the CMS of the flagged content determined that 37 of the posts constituted hate speech and 20 referred to false news, while five others incited violence and four constituted inauthentic behaviour.

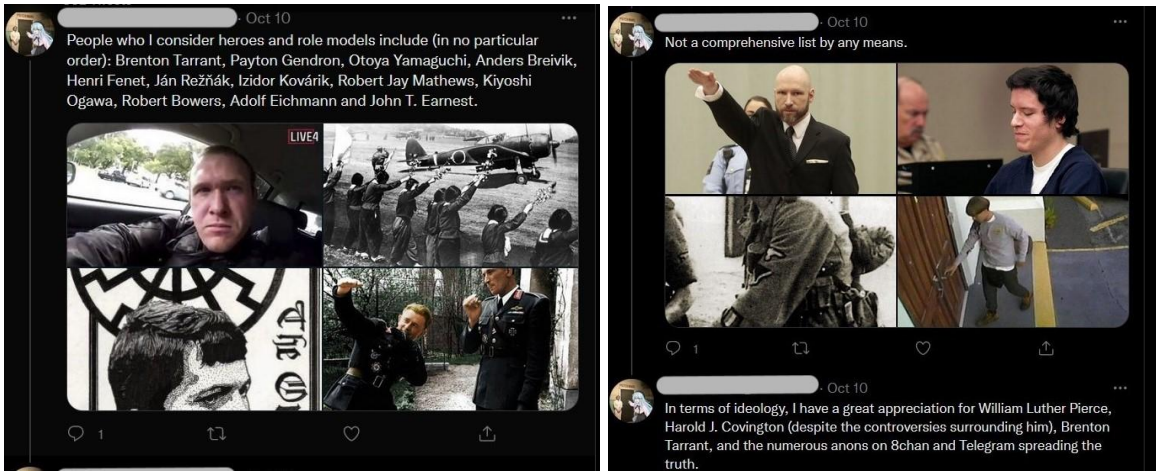
Example N: *An example of the type of reported content flagged for review on Facebook. English translation of the first comment: "If someone is gay, he should not annoy others with it. And these lgbti prides are just a provocation and I consider their behaviour there as gross indecency." English translation of the second comment: "These people should not be on the streets, they should be treated."*

The first piece of content flagged for removal by the CMS was reported to Facebook on 14 October (just two days after the attack). Yet by 21 November, Facebook had removed only six of the flagged posts with an average response rate of 11 days. In doing so, the platform sent just fifteen pieces of reported content to its third-party fact-checkers, with this action occurring sixteen days on average after the notification of problematic content violating the service's ToS was reported by the CMS. All told, Facebook failed to take any action in 70% of the total reported cases. The most frequent authors of content violating Facebook's ToS were notably repeat violators of Facebook's community guidelines who had been previously reported numerous times by the CMS before the terrorist attack took place. As of 3 January 2023 (82 days after the attack), Facebook had removed only 27 of the flagged posts, with an average response time of 39 days.

It is likely that Facebook's general failure to take any action on the vast majority of content flagged for review by the CMS can be attributed, at least in part, to the fact that the platform dedicates very few resources to third-party fact-checking review in smaller markets. In Slovakia, Facebook has contracted only one fact-checker for the entire country (from Agence France-Presse), and as a result knowingly delegates the fact-checking of all reported posts to a single contractor. The obvious absence of sufficient human resources unnecessarily prolongs the time during which problematic and harmful content is available online in Slovakia, as demonstrated in this case.

Notably, platforms also often interpret and apply their content moderation policies in a manner that decontextualizes a particularly problematic piece of flagged content and ignores the wider societal implications such content may carry. For instance, YouTube did not remove a video flagged by the CMS which, according to the national regulator's internal analysis, violated the platform's Hate speech and Violent criminal organisations policies. The platform replied, in a rather brief fashion, that the video in itself does not violate any of its policies. While platforms should not automatically remove content after receiving a notice from a regulator, they should however provide sufficient reasoning as to why they believe the reported content did not breach the platform's applicable policies or local laws.

As a platform directly involved in the dissemination of the Bratislava terrorist's problematic content, Twitter failed to enforce its own content moderation policies. Just two days before the attack occurred, the terrorist published the three tweets below, each of which feature illegal content such as far-right extremist symbols or praise of terrorist attacks in Christchurch or Utøya. These posts plainly violate Twitter's policies on hateful conduct and glorification of violence which prohibit the publication of symbols historically associated with hate groups and praise of events or actors that committed an act of violence against civilians.



Example O: *Obvious and clearly-recognizable Nazi symbols and pictures of numerous terrorists shared on Twitter by the shooter prior to the attack.*

Yet none of these tweets, or others dating back to April 2021 which also violated the platform's policies, were removed prior to the attack. And even though Twitter did act promptly to block the attacker's profile in the hours after the shooting took place, it was still too late to prevent his manifesto from circulating online after he posted it on the platform. According to information provided to the CMS by Twitter, after the attack the platform began blacklisting all links on Twitter leading to external sites containing the manifesto. Unfortunately, however, this action – while welcomed – did not cut off all access to the manifesto as the terrorist's entire Twitter account had already been uploaded to numerous online internet archives.

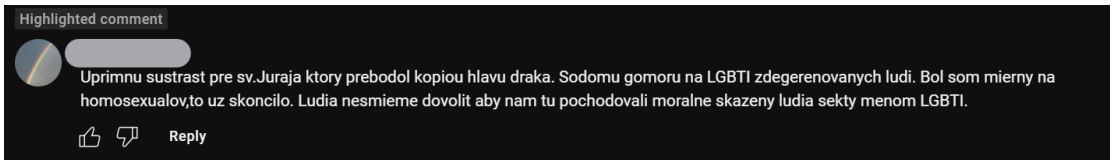
3. Evaluation of user comments after the attack

In response to the Bratislava shooting, researchers at Reset set out to investigate the prevalence of associated problematic and harmful content and assess platform behaviour in the aftermath of the attack. To do so, the team reviewed and analysed user comments posted under the posts with high interaction levels on Facebook, Instagram, and YouTube using keyword searches related to the event. The comments were all posted from 12 October to 23 October. Research results demonstrated that user debate occurring under posts reporting on the terrorist attack often included posts that violated platform ToS and, at times, Slovak law. For example, of the 300 most toxic comments identified through the research team's use of Perspective API,

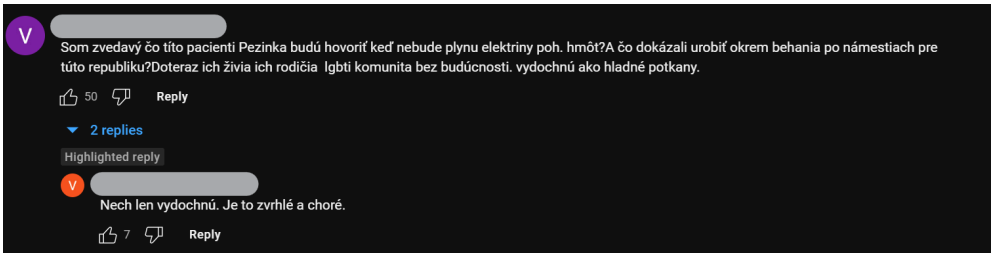
approximately every tenth comment used hate speech against the LGBTQ+ community that was exclusionary and dehumanising.

The research was conducted between 24 October and 5 November. Comments with derogatory and exclusionary language against the LGBTQ+ community were observed across all monitored platforms, with the share of comments deemed by Reset to constitute hate speech being comparable across all platforms (Facebook 10%, YouTube 10.6%, Instagram 12%). At the end of the investigation, researchers reviewed if the relevant content was still online and the results of the investigation were shared with the CMS.

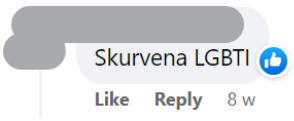
As shown in the examples below, problematic comments were found under a YouTube video of an interview with the attacker's father led by far-right politician Marian Kotleba. In the interview, the father of the attacker claimed – without evidence – that the attacker did not act consciously and was manipulated. The interviewer, in turn, cast doubt as to whether the shooter actually committed suicide – suggesting instead that he was executed. In response to these types of unsubstantiated conspiracy theories, one user commented by approving of the act of terrorism itself.



Example P: *A toxic comment posted under a YouTube video after the attack. English translation: "Sincere condolences for saint George (first name of the terrorist) who stabbed the dragon with his spear. Sodom and Gomorrah to LGBTI degenerates. I was moderate towards homosexuals, but that has ended now. People, we cannot allow these morally corrupt people from a sect called LGBTI to occupy public space"*.



Example Q: *A toxic comment posted under a YouTube video after the attack. English translation: "I am curious what these patients from a mental ward in Pezinok will say when there will be no gas or oil? What have they managed to achieve apart from running through squares for this country? To this moment, their parents feed them. LGBTI community without future. They die as hungry rats."*



Example R: *A toxic comment posted under a Facebook post after the attack. English translation: "Fucking LGBTI".*

V. Platform Cooperation with the CMS

a. Post-attack platform interactions

In the wake of the terrorist attack, the CMS engaged with Facebook, Twitter, and YouTube regarding the various measures the platforms had in place to mitigate the impact of terrorist and problematic content disseminated via their respective services. The results of this communication and cooperation are discussed below. As reflected, the differences in the content moderation policies of individual platforms appeared to render their case-by-case policy applications inconsistent and thus less effective overall given the everyday inter-service migration of users. Harmful and illegal content, like the Bratislava terrorist's manifesto, will continue to circulate and persist online as long as the dominant social media platforms fail to establish a common set of community standards vis-a-vis hate speech, disinformation, and other harmful content.

1. *Facebook*

The CMS has access to an escalation channel provided by Facebook (by far the most popular social media platform in Slovakia) which should give priority to content reported to the platform by the national regulatory authority. Using this channel, the CMS escalated a viral and potentially harmful video featuring disinformation related to the attack on 18 October, six days after the shooting occurred. Facebook promptly removed the video just a few hours after it was reported. The following month, the same video (although slightly edited) resurfaced on Facebook and was escalated by the CMS on 21 November. Facebook removed the video from the platform two days later. Both of these examples highlight the way by which platforms and regulatory authorities can work together in a positive manner to combat harmful content online connected to terrorism.

Yet despite these positives, there was an overall lack of responsiveness by Facebook. As highlighted above, Facebook failed to take any action in 70% of all reported cases by 21 November. Overall, despite a majority of the reported content being explicit and obvious hate speech, the platform seemingly attempted to avoid taking any strong action, such as content removal, and instead forwarded the potentially harmful and problematic pieces of content to its third-party fact-checker.

2. *Twitter*

Soon after the shooting, the CMS requested a meeting with Twitter representatives regarding the attack. During the 14 October bilateral meeting, Twitter representatives were notified that various archive iterations of the attacker's profile had been spread via Twitter despite the platform's blocking and blacklisting efforts. Pursuant to Media law no. 264/2022 Paragraph 110 (3)(q), the CMS requested access to the attacker's Twitter data. Despite clear legal competencies of the CMS to request information from digital platforms in this manner, Twitter

denied automatic access to the data and requested the CMS submit a lengthy law enforcement request to obtain it. Yet the CMS could not, however, go through the process and submit such a request as it had not been recognized by Twitter as a competent legal authority. When the form in question was finally able to be submitted on 8 November (almost a month after the CMS initially requested access to the attacker's data), the CMS had to wait until 24 November for Twitter to notify the national regulator that its law enforcement request had been rejected.

The communication between the CMS and Twitter in the wake of the shooting was overall slow and ineffective, reflecting poorly on the platform's content moderation commitments and practice. It is possible that the lack of proper communication and cooperation with the CMS could be attributed, at least in part, to Twitter's recent restructuring which resulted in hundreds of employees either resigning or being laid off.

3. *YouTube*

In the context of the Bratislava shooting, YouTube played a largely insignificant role in the dissemination of related far-right extremist content prior to or following the attack. The platform was, however, ineffective in its response to the aforementioned video interview of the terrorist's father that contained conspiracy theories regarding the event. The CMS notified YouTube of the problematic aspects of the video on the day of its publication (18 October), but the following day the platform declined to take down the video and referred to its own internal analysis stating that the video did not violate YouTube's community standards. By contrast, the same video was live-streamed on Facebook but then removed (due to its violation of Facebook's community standards) just a few hours after the CMS notified Meta of its existence. On 21 October, the CMS submitted an appeal to YouTube's decision not to remove the video from its service. Despite the appeal containing a detailed analysis of the video's contents and why it violated YouTube's ToS, the CMS received a response from YouTube on 25 October reiterating that the video did not violate the platform's community standards. Although YouTube did not take the action requested by the CMS, it did generally communicate in a prompt manner with the national regulatory authority regarding potentially harmful content.

b. Implications of the Digital Services Act

The forthcoming Digital Services Act (DSA), which will come into force in 2024, enshrines the principle that what is illegal offline is illegal online. This major shift in EU digital policy strives to tackle illegal content, online discrimination, and other potentially harmful online behaviour. The DSA takes a tiered approach in its scope and will to a certain extent hold very large online platforms (VLOPs), including the major social media platforms addressed in this report, liable for the content available through their services.

In working to address the persistent shortcomings of VLOPs vis-a-vis their services enabling potential and actual societal harms, the DSA sets forth the following obligations:

- Harmonisation of processes through which platforms are notified of illegal content and the actions they ought to take to expeditiously remove such content (Article 16). The DSA further allows the Digital Services Coordinator (DSC) in each member state to designate the so-called trusted flaggers – civil society entities that have demonstrated sufficient expertise in a particular field – whose reports of illegal content ought to be treated with priority (Article 22).

- Creation of risk assessment and risk mitigation measures that require VLOPs to assess systemic risks and put in place effective measures to address them (Articles 34 and 35). With regard to content moderation, the DSA encourages platforms to address the speed and quality of processing notices related to specific types of illegal content, especially illegal speech, and their subsequent removal.

- Each EU member state is to appoint a DSC that will be responsible for the enforcement of the DSA at the national level (Art 49). This creates a network of DSCs able to tackle EU-wide problems, such as acts of terrorism, without due delay. In addition to its coordination role, the DSC will serve as the complaints body for all users and will vet those researchers seeking access to platform data. The DSC will also be responsible for drafting annual reports which will include the number and subject matter of orders to act against illegal content and the effects given to those orders.

The effective enforcement of the DSA has the potential to significantly enhance the contemporary legal frameworks, such as the EU regulation on countering terrorist content online (TCO), for tackling harmful and illegal content. For example, the DSA bolsters the capacities of the individual member states to issue notifications both regarding terrorist content to hosting service providers and hate speech to VLOPs. Had the DSA been in place before the attack, the monitored VLOPs, especially Facebook and Twitter, would have had clearer and enforceable legal obligations to react quickly against potentially illegal and harmful content after the attack and react more promptly to requests of CMS. In the future, platforms will need to set up systems and mechanisms to assess the risks of their services, within which they should asses also the societal impact, such as the proposed baseline standards for terrorist and extremist content, which would have prevented the publication and spread of such content. Moreover, the DSA is to streamline the communication process between the competent authorities and the respective platforms, which could have contributed to faster takedown of potentially harmful or illegal content. Overall, the requirements set out in the DSA for the VLOPs would have facilitated a more effective cooperation between the CMS and the platforms covered by this report.

VI. Conclusion & Recommendations

The October 2022 terrorist attack in Bratislava once again illuminated the outsized role social media platforms play in the online dissemination of illegal and harmful content. In cooperation with the CMS, Reset identified numerous shortcomings in platform enforcement of content moderation policies and demonstrated how platforms often react slowly to remove illegal and harmful content – even when notified of its problematic nature by a national regulatory authority. The inability of large-scale digital platforms to enforce their own policies is especially pronounced in smaller markets like Slovakia, where the resources devoted to content moderation and human review are minimal. The results also reveal the inability of major platforms like Twitter to monitor and identify extremist, terrorist, and hateful content – even when the language used in such content is English.

Having reviewed the respective responses of Facebook, Twitter, and YouTube to a crisis situation in Slovakia, it is clear that more consistent and enforceable regulatory oversight is needed. Such reform will hopefully be coming, at least in part, via the DSA under which certain platforms will be required to begin setting up systems and mechanisms to assess the risks their services pose to European citizens. In addition to complying with this landmark new piece of European legislation, authors of this report encourage platforms to work together to create a common baseline for the definition and identification of extremist and terrorist content. As evidenced in the Bratislava case, subtle differences in platform policies, and/or narrow interpretations thereof, can combine to foster significant obstacles for those seeking to promptly and efficiently prevent problematic content from circulating online.

Considering the recent rise in lone-actor terrorist attacks, platforms should also engage with experts from academia and local NGOs to develop country-specific methodologies for the assessment of terrorist content online. This is particularly important given that the status quo, which is often based on prescribed lists of terrorist organisations, has proven to be ineffective and out-of-step with current terrorist activity and trends. Efforts must be made to account for the fact that an increasing number of attacks are carried out by lone militant accelerationists and individual actors.