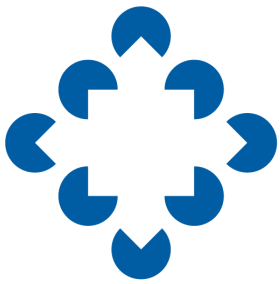


The prevalence of harmful or potentially illegal content on digital platforms following the Bratislava terrorist attack

Council for Media Services
& Trust Lab

Key Findings

The Council for Media Services (CMS) conducted a study, in liaison with Trust Lab, on the prevalence of harmful and potentially illegal content available on four major digital platforms (Facebook, Instagram, Youtube and TikTok) following the Bratislava terrorist attack. The study introduces its data collection methodology and legal framework for assessing the legality of online content and interprets the results in light of the AVMSD and DSA. The findings point out the failures of the platforms to moderate content and hint at an emerging threat in the form of borderline content.

Prevalence of harmful or potentially illegal content online: The analyzed platforms still host harmful and/or potentially illegal content related to the terrorist attack. Access to such content is relatively easy as Trust Lab's monitoring identified **253 unique links**, of which 123 were posts, and 130 were comments.

Content moderation failures: Trust Lab reported all 253 instances of harmful and/or potentially illegal content to the respective services via their user reporting mechanism. Despite these reports, the platforms removed only **12** pieces of content (Facebook - 8x, Youtube - 3x, Instagram - 1x). TikTok did not take action against any of the reported content.

Non-functional reporting mechanism: Considering the absence of any meaningful response from the platforms, it may be concluded that the mechanisms for reporting harmful and/or potentially illegal content available to users are non-functional. It is only after the problematic content was reported by a national regulatory authority that the platforms reacted swiftly and removed all **26 instances** of content reported by the CMS.

Borderline content: The monitoring revealed several instances of borderline content, which is a type of content that falls just short of violating the Terms of Service (ToS) or Union/national law. However, this type of content still contains misleading, harmful or potentially illegal elements, such as incitement to violence or praise of terrorist acts. Its ambiguity, however, precludes both the regulator and the platforms from enforcing the law and ToS respectively.

Introduction

The Council for Media Services is the Slovak national regulatory authority instituted by Act no. 264/2022 of the Slovak Republic (The Media Services Act - MSA). The CMS is responsible for media oversight and enforcement of regulatory frameworks concerning retransmission, broadcasting, provision of on-demand audiovisual media services, and online platforms. As part of the CMS's mission to exercise state regulation within the digital domain, the regulator tackles harmful and potentially illegal content which poses a significant risk to the general public.

As a consequence of the terrorist attack on Zámocká Street in Bratislava, there was a surge in harmful and potentially illegal content available online.[1] To monitor the availability of such content on online platforms, the CMS commissioned, in November 2022, Trust Lab to conduct a thorough monitoring of the situation, as the CMS does not yet have the technological means to collect large quantities of data across different platforms, especially with regard to the types of content analyzed. As such, Trust Lab's monitoring tackles the lack of access to the platforms' data by using an outside-in approach that is independent of any API or backend data access, which significantly exacerbates the enforcement of applicable national or Union law. The monitoring covered harmful or potentially illegal content related to the terrorist attack on four platforms, namely Facebook, Instagram, Youtube, and TikTok. The data for this study were collected between December 5 and December 9, 2022. Specifically, the monitoring focused on the findability of content featuring hate speech, harassment based on protected categories, or violent extremism. The searches included keywords related to the Bratislava terrorist attack and the subsequent wave of hate speech against the LGBTI+ community. Trust Lab's proven and vetted approach to assessing the prevalence of harmful online content can help regulators, public institutions, and private companies better understand the prevalence and spread of such harmful content across different platforms and markets, thus allowing them to deploy new strategies, accurate policies, and improved enforcement to tackle the harmful content and thus protect their users.

The following chapters thus detail the methodology for collecting data from the platforms as well as the legal framework used for the assessment of legality. Following the analysis, the results are interpreted in light of the already transposed Audiovisual Media Services Directive (AVMSD) and the upcoming[2] Digital Services Act (DSA). Finally, the regulator puts forward a list of recommendations for the platforms to further their efforts in tackling harmful and potentially illegal content online.

[1] Read more about the attack in the reports published by CMS and in cooperation with Reset.
[2] as of April 2022

Methodology

Trust Lab conducted a comprehensive monitoring of content featuring hate speech on social media platforms in Slovakia, with a special focus on hate speech related to the terrorist attack on Zámocká Street in Bratislava. The study measures the findability of such content and the frequency and timing of social media platforms' actions on that content. The study includes four platforms, namely Facebook, Instagram, Youtube, and TikTok. The comparison across markets and platforms will enable the ranking of individual platforms based on their performance, among others, to understand which platform has the highest exposure to content related to the Zámocká Street attack.

The monitoring utilized Trust Lab's patent-pending Kaptix technology, which captures three content-related metrics: findability, removal rate, and time-to-action. Findability measures the volume of "true positive" social media entities discoverable by a motivated searcher over a specific time period. The removal rate monitors the content found over time, noting any status changes such as warning labels, age restrictions, or content removals. Time-to-action calculates the time a platform takes to take an eventual action. The combination of the three metrics provides a statistically significant perspective on a social media platform's performance. These metrics provide answers to three intertwined questions:

1. How much controversial or compromising content can be found by a vulnerable and/or inquisitive user?
2. How strictly a platform reacts to that content?
3. How operationally effective a platform is in carrying out that action expeditiously?

For this monitoring, Trust Lab spent 120 hours looking for Slovak content related to the terrorist attack on Zámocká Street on four social media platforms (Facebook, Instagram, Youtube and TikTok) and found 253 unique links of harmful and potentially illegal content. For the searches, Trust Lab used in-market native speakers with knowledge of local customs, culture, and political situation. The selection of the keywords and platforms was randomized to minimize learning bias.[3] All social media content identified by the searchers was audited for quality and coded by both Trust Lab policy experts and the CMS' analytical department.

[3] Keywords: #Teplaren, #Vražda na zámockej, #Zamocka, #Teroristický útok v Bratislave, #Lgbti slovakia, #Lgbti, #lgbti komunita, #Strelec zo Zámockej, #Strelba lgbt, #Útok na LGBTI+ komunitu.

Slovak legal framework for assessing potentially illegal content

The competences of the CMS as the Slovak national regulatory authority responsible for media oversight and enforcement of regulatory frameworks concerning potentially illegal content on digital platforms are instituted by the MSA.

Under MSA, the CMS is entrusted with the responsibility and legal competency to prevent the dissemination of illegal content online. This entails cooperation with the platforms in the effective, proportionate, and non-discriminatory application of their community rules, norms, and standards as per Article 110 (3)(q) of MSA. In this regard, Article 151 (2) of MSA 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 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.

Results

Monitoring

From the 253 links containing harmful or potentially illegal content, 123 were posts and 130 were comments. All 253 links were reported by Trust Lab using the platforms' user interfaces, which are designed to allow users to report content that violates their Terms of Service (ToS). As a result, the platforms took action in 4.7 % of cases.

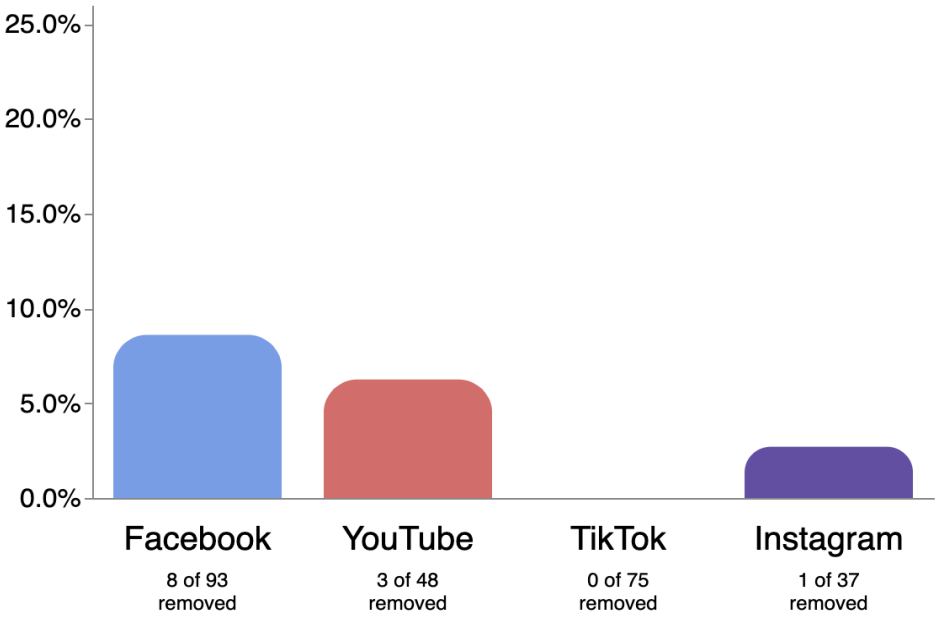


Figure 1 - The number of posts or comments removed by the platforms following a report by Trust Lab

Even though action against objectionable content was taken, it took Facebook and Youtube 32 days to remove the content, eight and three pieces respectively, reported by Trust Lab. Besides the monitored metrics, Trust Lab also found that the vast majority of videos, often containing harmful or potentially illegal content, were monetised. Moreover, the contents of comments under the analysed posts featured harmful or potentially illegal language that by far surpassed, in terms of the perceived level of harm, the posts themselves.

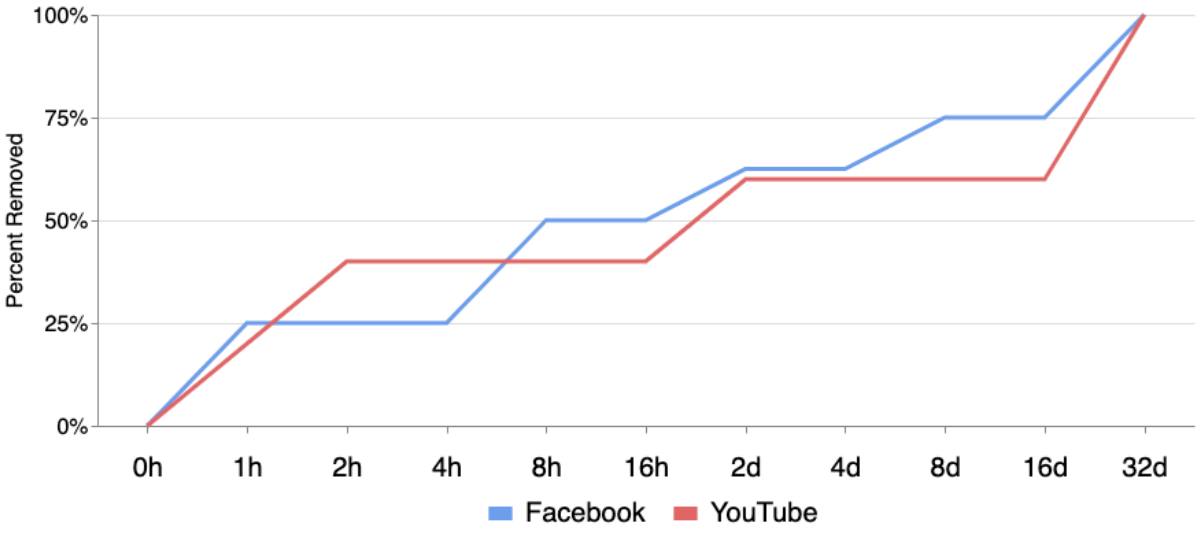


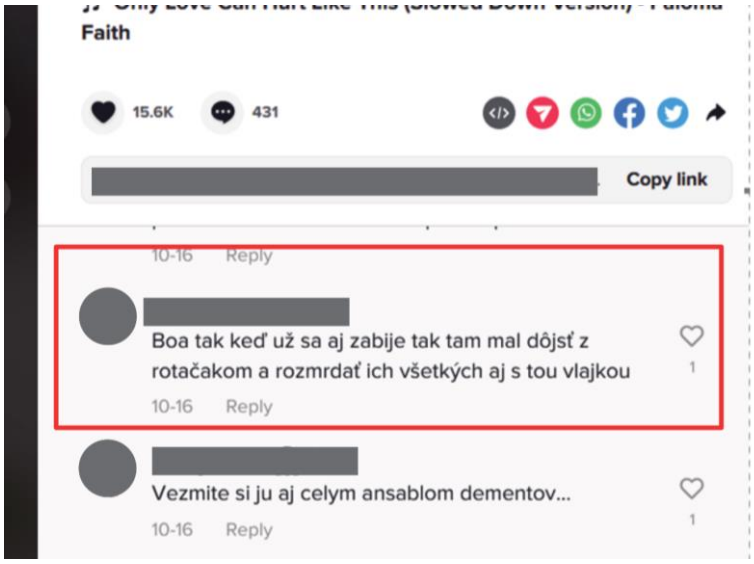
Figure 2 - The number of days it took for Facebook and YouTube to remove eight and three, respectively, pieces of content reported by Trust Lab

Analysis of content

The volume of content identified by Trust Lab was analysed using a qualitative content analysis (including both latent and explicit meaning) method. The analysis investigated the contents of posts and comments published by users on the four platforms while considering the contextual nuances surrounding the case at hand. Following the analysis, the CMS reported the content it deemed as potentially illegal to the platforms using dedicated escalation channels.

The CMS reviewed all 253 reported posts and comments identified by Trust Lab. After a closer examination of the intensity and contextual setting of the content, the CMS identified 26 posts and comments on three services (17x TikTok – ByteDance Ltd., 8x YouTube – Google Ireland Limited, 1x Facebook - Meta Platforms, Inc.) that might have potentially constituted illegal content according to the article 151 (2) of MSA.

The posts and comments analysed by the CMS had been published by platform users as a response to the high coverage of the terrorist attack in the media (news about developments in the police investigation, relevant contextual information concerning the attack etc.). Common features of the posts were: the disputing of the attacker's motives and explicit approval of the Bratislava terrorist attack on the LGBTI+ community as well as the incitement of violence against the LGBTI+ community.

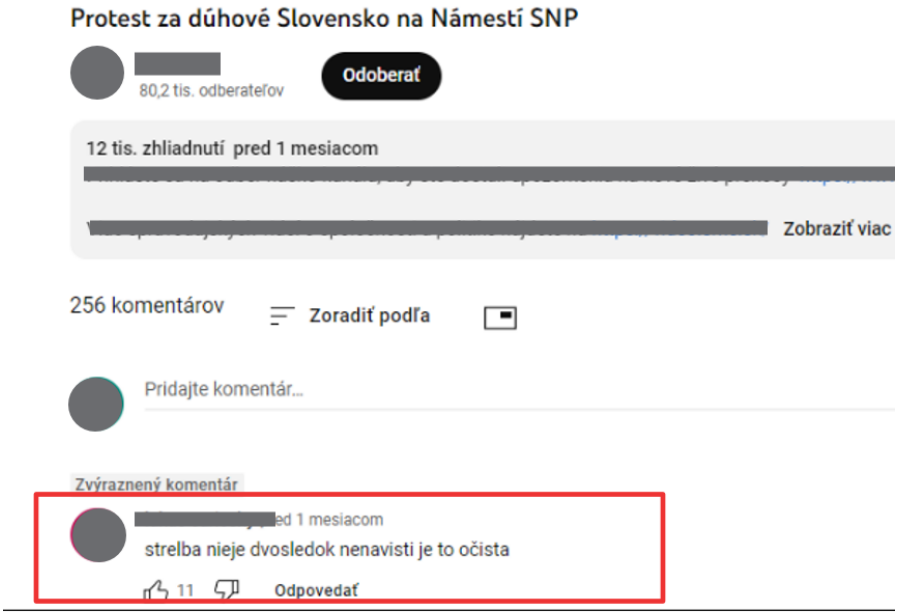


Screenshot 1 - Potentially illegal content on Tiktok

Text of comment: *"Boa tak keď už sa aj zabije tak tam mal dôjsť z rotačakom a rozmrdať ich všetkých aj s tou vlajkou"*

Eng: *"God so if he kills himself he should have come there with a machine gun and smashed them all with the flag as well"*

Analysed posts and comments either praised the attacker for shooting members of the LGBTI+ community or expressed their pity that there were only two victims. The content at hand further claimed that the attack was provoked and the shooting was an act of self-defence in both literal and figurative meaning. Similarly, the content included claims regarding either the need to segregate the LGBTI+ community from other citizens or banish them from Slovak Republic altogether.



Screenshot 2 - Potentially illegal content on YouTube

Text: *"strelba nieje dvosledok nenavisti je to očista"*

Eng: *"shooting is not the result of hatred, it is a cleansing"*

The analysis concluded that the posts and comments constitute a type of content that incites violence or hatred against a group of people based on a protected category, it being their real or assumed sexual orientation. Such content further calls for the restriction of their rights and freedoms as per article 424 (1) of Act No. 300/2005, Penal Code.[4] Therefore, the CMS took action as per article 110 (3)(q) CMS and contacted the providers of the platforms (Tiktok, Youtube and Facebook), thereby notifying them of potentially illegal content on the services.

To date (9th March 2023), ByteDance Ltd. replied that all 17 pieces of reported content were removed from TikTok. Google Ireland Limited removed all 8 pieces of the reported content. Finally, Meta Platforms, Inc. notified the regulator that it had removed the

[4] An article 424 (1) of Act No. 300/2005, Penal Code: "Whoever publicly incites violence or hatred against a group of persons or an individual because of their real or assumed membership of a race, nation, nationality, ethnic group, because of their real or assumed origin, colour, sexual orientation, religion or because they are non-religious, or publicly incites the restriction of their rights and freedoms, shall be punished by imprisonment for up to three years."

reported content. All platforms claim that the reported content violated their respective Community Guidelines and ToS.

Borderline Content

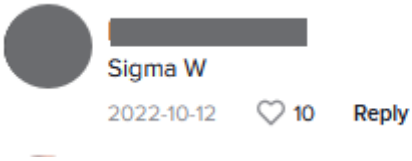
While content violating the platforms' ToS or Union/state law may be clearly identified based on transparent metrics, borderline content constitutes a more elusive phenomenon. Borderline content may be defined either as legal, insofar as freedom of speech goes, but problematic content not appropriate for the public[5] or content that is on the border of violating a platform's policy.[6] Although often overlooked in the reporting on the prevalence of harmful content, the CMS identified, based on Trust Lab's dataset, a number of comments, predominantly on TikTok, that constitute borderline content.

The identified borderline content praised and glorified the perpetrator for carrying out the terrorist attack. The praise, albeit implicit, shows support for the attacker by using coded language native for younger users of social media and fringe forums like 4 or 8chan.

Screenshot 3 - Borderline content on TikTok

Text: *"až ho chytnout tohle mu vrátim"*

Eng: *"once they catch him, i will give him this back"*



Screenshot 4 - Borderline content on TikTok

Text: *"Sigma W"*

Meaning: Sigma refers to a pseudo-scientific construct of the masculinist alt-right denoting a particular type of male behaviour (successful, highly independent, intelligent). A standalone letter "W" in this context means "a winner" which, in contemporary slang, denotes a congratulation to someone's success. In

[5] Heldt, A. (2020). Borderline Speech: Caught in a Free Speech Limbo?. *Internet Policy Review*, Op-Ed.

[6] Gillespie,T. (2022). Reduction / Borderline content / Shadowbanning. In *Platform Governance Terminologies*. *Yale Journal of Law and Technology*. Series.

this context, the comment praises and glorifies the perpetrator for carrying out the attack.

In both cases, the content becomes potentially illegal only after careful analysis and assessment (considering that one has the knowledge to do so). Moreover, the content utilizes emojis, symbols and slang that are increasingly difficult to interpret due to the fast-paced and changing nature of online discourse. As already noted in the CMS' latest report, even the state-of-the-art content moderation systems deployed by the largest platforms are unable to detect borderline content, which in turn contributes to the spread of harmful content and potential radicalization.

As such, borderline content represents a formidable challenge to both regulators and social media platforms. For it is the ambiguity and vagueness of borderline content that the CMS is not able to utilize its enforcement powers and request a swift removal of such content. The amplification of borderline content coupled with the platforms' inaction may, however, constitute a systemic risk under the upcoming DSA regime (article 34). Therefore, we believe that more attention should be given to these subtle but harmful means of communication that have the potential to contribute to offline harms.

Relevant legislation

The prevalence of harmful and potentially illegal content on the largest social media platforms following the Bratislava terrorist attack is of no surprise, as previous reports published by CMS and in cooperation with Reset already note. While the previous reports highlight failures in content moderation, this study provides a new, arguably quantitative, perspective using Trust Lab's Kaptix technology to analyze a larger data corpus (N = 253). The discussion on online societal harms is inextricably linked to a range of governmental policy upgrades. For this, the results are interpreted in light of the Audiovisual Media Services Directive (AVMSD) and the forthcoming Digital Services Act (DSA).

Audiovisual Media Services Directive (AVMSD)

Audiovisual Media Services Directive (AVMSD) sets an EU-wide minimum harmonised legal framework in the audiovisual internal market. During its last major revision in 2018, its scope was enlarged beyond editorially responsible AV media services (TV and on-demand) and also to video-sharing platforms (VSPs). VSPs are uniquely defined in the AVMSD and represent only a subset of the online platforms under the DSA regime. AVMSD is a sectorial legislation which primarily addresses audiovisual content. As a directive, it is governed by a country of origin principle, meaning its rules are enforced only in the member state where the service is established.

Application still in progress

Unfortunately, the transposition of the AVMSD has been delayed across the EU. Although the AVMSD has been transposed into the Slovak legislation through the adoption of the MSA a few months before the attack, none of the mentioned platforms was registered in Slovakia as per Article 186 of the MSA. Moreover, no platform was, neither by the time of the attack nor at the time of writing of this report, officially registered as a VSP per the Mavise database (for now the mentioned services are included only informally without a designated regulator). In practice, therefore, the procedures (including when dealing with cross-border situations) and rules contained in the AVMSD and in the MSA could not be applied. As a result, the effectiveness of the rules applicable as per the relevant AVMSD articles could not be assessed in the case of the monitored services.

Sectoral assessment of measures

In a similar fashion to the DSA, AVMSD is focused on systematic oversight with the role of the regulator mainly to assess the appropriateness of the measures implemented by

the VSPs as specified in the AVMSD.[7] The rationale behind the Directive is to protect the general public, with a special emphasis on the protection of minors, from harmful and/or potentially illegal content. The findings above hint at concerns on the systematic level regarding the appropriateness of the measures applied by some of the monitored services in this report as it relates to the obligations from the AVMSD (e.g. establishing and operating transparent and user-friendly mechanisms for users of VSPs to report or flag content or easy to use and effective procedures for the handling and resolution of users complaints).

Role of CMS

This report and some of the previous work of the CMS will therefore be crucial once the system is fully set up with the designated VSPs. It will be crucial to reach out formally and informally and feed into the work of the responsible regulator in the country of origin of the VSPs. In the agenda of international cooperation, the CMS cooperates intensively with other regulatory authorities that will have oversight and enforcement powers over the largest VSPs. In this regard, the CMS has the legal competence to cooperate with other competent authorities, including in the area of VSPs.

[7] Article 28b AVMSD: "take appropriate measures to protect: (a) minors from programmes, user-generated videos and audiovisual commercial communications which may impair their physical, mental or moral development in accordance with Article 6a(1); (b) the general public from programmes, user-generated videos and audiovisual commercial communications containing incitement to violence or hatred directed against a group of persons or a member of a group based on any of the grounds referred to in Article 21 of the Charter; (c) the general public from programmes, user-generated videos and audiovisual commercial communications containing content the dissemination of which constitutes an activity which is a criminal offence under Union law, namely public provocation to commit a terrorist offence as set out in Article 5 of Directive (EU) 2017/541, offences concerning child pornography as set out in Article 5(4) of Directive 2011/93/EU of the European Parliament and of the Council (1) and offences concerning racism and xenophobia as set out in Article 1 of Framework Decision 2008/913/JHA."

Digital Services Act (DSA)

Coming into force in 2024, the DSA seeks to build a safe online environment effectively tackling illegal content, risks posed by very large online platforms and search engines (VLOPSEs) and online discrimination. The DSA takes a tiered approach in its scope which is reflected in the proportional assigned of obligations to intermediary services. This study includes the four largest platforms operating on the EU market[8] and thus interprets the results in light of their obligations.

Risk assessment

One of the most demanding obligations of the DSA is for VLOPs to assess the societal risks posed by their respective services (article 34). Purposely, the definition of systemic risks is broad but includes, for this study crucial, the dissemination of illegal content and any actual or foreseeable negative effects in relation to gender-based violence. When conducting the assessment, the platforms ought to evaluate the effectiveness of their content moderation systems, enforcement of their ToS, and the design of their recommender systems. Upon assessment, the platforms are required to adopt reasonable risk mitigation measures tackling the deficiencies identified in the assessment.

The findings demonstrate that the largest platforms pose significant risks to society, especially regarding social polarisation and incitement to violence. The identified content highlights the failures of the platforms to limit the spread of harmful and manifestly illegal content following a terrorist attack. Such content may not only incite further violence but also radicalise other users. Thus, it seems the platforms have not established effective mechanisms to prevent the spread of problematic content on their services following a crisis.

Notice and Action mechanisms

Under the DSA regime (article 16), all platforms must provide users with a mechanism that allows them to inform the platforms of illegal content. The notices issued using such mechanisms should constitute notifications of illegal content, which prompts the platform to act swiftly with regard to the assessment and possible removal of the content in question.

[8] According to the transparency reports issued by the individual companies, all four platforms will be most likely designated as VLOPs according to article 33 of the DSA.

The results of Trust Lab's reporting activity show that currently available mechanisms through which users report problematic content are not effective. For instance, TikTok did not react to any of the reports made by Trust Lab. The potentially illegal content identified by the CMS was removed only after a national regulatory authority notified the platforms via dedicated escalation channels. The slow and inefficient platform responses to user reports of potentially illegal content may thus eventually constitute an infringement making the platforms liable for the nature of the content available through their services.

Trusted flaggers

The status of a trusted flagger has already been recognised by numerous platforms (e.g. Youtube). The DSA harmonises the requirements for becoming a trusted flagger and allows trusted flaggers to monitor any platform operating in the European market (article 22). Considering the importance of Notice and Action mechanisms, trusted flaggers are providing the platforms with the expertise that is needed for a swift assessment of legality. The DSA requires the platforms to prioritise the flaggers' reports and process them without undue delay.

As mentioned above, the platforms seem to assess and remove content only after a national regulatory authority has flagged it. It is believed that trusted flaggers will help platforms process more reports at a faster pace while retaining independence and providing expertise on the matter at hand. The findings of this study only highlight the need for trusted flaggers in regions that have been so far neglected by the platforms with regard to content moderation.

Access to data

Currently, access to VLOPs' data is limited which precludes both proper public scrutiny as well as any research into online harms facilitated by the platforms. The upcoming DSA requires platforms to provide data access to vetted researchers who will be thus able to contribute to the detection, identification and understanding of systemic risk as well as to the assessment of the adequacy of the risk mitigation measures (article 40).

The results demonstrate the difficulty of accessing platform data, which requires the use of complex, patent-pending, technologies such as Trust Lab's Kaptix. This obscurity prevents effective oversight over potentially illegal content as well as any other hate-speech-related harms, such as incitements to violence. Therefore, it is necessary that Digital Service Coordinators (DSCs) oversee the effective implementation of art. 40 of the DSA in order to streamline the access to platforms' data.

Recommendations

The findings of this study highlight the prevalence of harmful and potentially illegal content on the four largest digital platforms in the months following the terrorist attack on Zámocká Street. It is clear that the platforms have failed to moderate problematic content despite having been notified of it by their users. The following section thus puts forward a number of recommendations to the platforms as well as hints upon the future venues for the research into online harms.

User reporting

In the process of examining the platforms and their actions in moderating content, it was found that **the processes for reporting problematic content to the platform by users often lack transparency and effectiveness**. While all platforms allow users to report problematic content, users are often unable to review their reports and platforms provide little transparency regarding their assessment of the reported content. Furthermore, the content reporting mechanisms deployed by the analysed platforms all reveal severe neglect of the users' reports. Put differently, platforms are highly unlikely to act upon a piece of content that was reported by an average user. To tackle this, platforms ought to refine the reporting process, provide sufficient transparency and dedicate enough resources in order to review content reported as harmful or potentially illegal. More human and financial resources should be allocated to small and medium-sized markets to remove systematic barriers, such as the adequate size of content moderation teams and fact-checking teams. Regarding illegal content, to comply with the upcoming DSA, platforms ought to prioritize notice and action mechanisms to allow individuals or entities to notify them of its presence. These mechanisms must be easy to access and user-friendly.

User interface

The review of the collected data reveals barriers for researchers as well as national regulatory authorities in the pursuit of monitoring harmful or potentially illegal content. The user interface of the analyzed platforms does not allow for data portability with regard to URLs. For example, Instagram does not allow users to get the URL of individual comments posted under a post on the user's feed. In contrast, TikTok does not allow users to use the 'Find Command' (the Ctrl+F shortcut) to search for content in the comments section. These barriers, albeit minor, effectively preclude any research or monitoring of potentially illegal content and thus internet harms.

In order to remedy the situation, platforms ought to take steps to provide users, researchers and regulatory authorities with interfaces that foster transparency and

promote harm mitigation. All content accessible via the main user interface should be thus shareable outside of the platform.

Policy review and borderline content

The findings of this study showcase a number of phenomena that occur on platforms and contribute to potential offline harm. One such phenomenon is the “gaming the system” effect. As soon as (malicious) users discover that certain types of content violate the ToS or could lead to other repercussions, they alter their communication so as not to be penalised. These changes often involve slang or the use of symbols and emoticons. For example, in the case of problematic posts related to the attack on Zámocká Street, expressions such as "juraj W" or "he deserves [👑]" were used. Users expressing their support for acts of terrorism in this fashion effectively avoid detection by the platforms and preclude national regulatory authorities from enforcing state or the Union law as a result of the messages’ ambiguity.

This and other similar content may be referred to as borderline content, which is content that falls just short of violating platforms’ ToS, but still contains potentially harmful or controversial elements. It can include misleading, inflammatory or divisive content and can sometimes cross the line between free speech and harmful conduct.

Effectively regulating borderline content on platforms requires a multi-faceted approach that combines technological solutions, human judgment, stakeholder engagement, transparency and accountability. By adopting these strategies, social media platforms can more effectively address the challenges posed by borderline content and promote a safer and more responsible online environment.

This approach can be divided into four main points:
1.) Comprehensive and clear ToS: Platforms’ policies should define what constitutes borderline content and how to deal with it. Policies should be effectively communicated and reflect the communication style of users - especially in cases of potentially illegal content.

2.) Content moderation: Train AI-powered content moderation tools to identify defined borderline content in order to take action more swiftly: use these technologies to identify problematic content and take action more quickly.

3.) Increase the number of moderators and collaboration with experts: Platforms should dedicate more resources to their content moderation efforts by, for example, hiring more moderators or collaborating with experts on harm mitigation. The lack of resources is most evident in minor markets and minority languages.

4.) Transparency and accountability: Platforms should be more transparent about their ToS and enforcement policies. They should be accountable to their users and allow for feedback and processes to ensure that enforcement decisions are fair and consistent. They should also report regularly on their progress in dealing with borderline content.

Downranking of content following a crisis situation

In the wake of a crisis, information channels are usually subject to a barrage of mis/disinformation. In the case of terrorist attacks, not even the state officials on the ground usually have the necessary information. At the same time, it is commonly observed that there can be a proliferation of hate speech in the aftermath of a terrorist attack. This phenomenon was well-documented after the attacks on mosques in Oslo and Christchurch. Studies[9] have shown that there was a significant rise in anti-Muslim sentiment and Islamophobia. Similarly, after the terrorist attack at Zámocká Street, there has been a significant increase in hateful and mis/disinformation content on social media.

Therefore, online platforms should provide the state authorities with emergency mechanisms allowing state officials to inform the platforms of a crisis scenario. Using real-time content insights, platforms should downrank content related to the crisis in its first 48 hours. Specifically, platforms may cooperate with state authorities, experts or researchers to develop a list of keywords or phrases for each language that is likely to be associated with harmful or potentially illegal content related to the crisis. In the case of a terrorist attack, these keywords may include the name of the attacker or attackers, the location of the attack, the names of other terrorists or extremist terrorist groups, and specific words associated with hateful speech against the targets. Downranking content should be done in a transparent, consistent, and fair way, and platforms should communicate their policies and practices to users and provide opportunities for feedback. Additionally, platforms should prioritize informative and accurate content delivered by state officials and state institutions´ pages, especially during the first hours after the attack.

[9] https://cepr.org/voxeu/columns/jihadi-attacks-media-and-local-anti-muslim-hate-crime; https://s3.documentcloud.org/documents/21416310/islamophobia-report-3-2022-hr-pages-ra.pdf

About the Council for Media Services:

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 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.

The CMS has the 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. Besides, the CMS has the competency to assess the appropriateness of public protection measures adopted by the platforms within the Slovak market.

Among its other activities, the CMS is an active member of well-renowned international platforms for regulatory authorities, such as ERGA (European Regulators Group for Audiovisual Media Services) and EPRA (European Platform of Regulatory Authorities), as well as international initiatives tackling terrorism and other types of illegal content online, such as the Christchurch call and GIFCT (Global Internet Forum to Counter Terrorism).

About Trust Lab:

Trust & Safety Laboratory was founded in 2019 by senior Trust & Safety leaders from Google, YouTube, Reddit, and TikTok with a mission to make the web safer for everyone. As leading executives responsible for Trust & Safety Engineering, Product, Ops and safety engineering, product operations, and policy for over a decade each, they build large-scale systems and tools to identify harmful, policy-violating, or otherwise unsafe content, accounts, and transactions on online platforms. Trust Lab is a data science and technology company with headquarters in Berlin, Germany, and Palo Alto, California, USA. It also has offices in other places around the world and attracts international talent.

Trust Lab works with a broad spectrum of clients in the private and public sectors and partners with stakeholders across the online safety industry. They offer measurement and moderation systems, tools, and services for harmful online content, with a particular focus on disinformation and misinformation. Among our clients are many

leading social media platforms, marketplaces, the European Commission, and the US Government. They have measured a wide range of online harms including the prevalence of misinformation related to the coronavirus pandemic, the prevalence of hate speech, and we are currently running a 40-week project with DG HOME on the measurement of algorithmic amplification as well as the prevalence of Terrorism and Violent Extremism content across number of European languages. Trust Lab delivers one-off insights or continuous reporting on a monthly basis, presented via an interactive dashboard, reports and visualisations in presentation form. In the context of the upcoming implementation of DSA, they build large-scale systems and tools to help companies with monitoring, detection and regulatory compliance of harmful, illegal or otherwise unsafe content and accounts on their platform.