OECD publishing

TRANSPARENCY REPORTING ON TERRORIST AND VIOLENT EXTREMIST CONTENT ONLINE

AN UPDATE ON THE GLOBAL TOP 50 CONTENT SHARING SERVICES

OECD DIGITAL ECONOMY PAPERS
July 2021 **No. 313**

OECD
BETTER POLICIES FOR BETTER LIVES

Foreword

Following its 79th Session in July 2019, the Committee prioritised its preferences for follow-on work related to online platforms, based on the options presented in document DSTI/CDEP(2019)7/REV1. A project to develop a voluntary transparency reporting framework and metrics for terrorist and violent extremist content (TVEC) online was one of two projects to be undertaken in the current Programme of Work and Budget, the other being a study of data portability (see DSTI/CDEP/DGP(2019)2). Stage One of the TVEC project calls for two reports, spaced one year apart, that take stock of the current policies and procedures related to TVEC of the world's leading online platforms and other online content sharing services. The first report, Current Approaches to Terrorist and Violent Extremist Content among the Global Top 50 Online Content Sharing Services, was published in 2020. This final draft of the second report takes into account the oral and written feedback from delegates on the first and second drafts, as well as feedback from the profiled companies in Annex B.

The TVEC project is proceeding with the kind financial support of Australia, Korea and New Zealand. The Secretariat would like to thank Dr Tomas Llanos for his work on this report.

Note to Delegations:

This document is also available on O.N.E under the reference code:

DSTI/CDEP(2020)9/FINAL

Table of contents

Executive Summary

Terrorist and violent extremist groups use the Internet and associated technologies for radicalisation, recruitment, dissemination of propaganda, communication and mobilisation. Terrorist and violent extremist content (TVEC) posted online can be disseminated quickly and cheaply, amplifying dangerous views and reaching broad audiences. This is a second report examining the TVEC-related policies and procedures of the world's top 50 online content-sharing services. The first report provided a benchmark against which this second report assesses relevant developments, such as whether more or fewer services publish transparency reports on TVEC.

As the experience of the Christchurch and Halle attacks demonstrated, gruesome and shocking acts of violence can be broadcast unedited and in real time online. Such tragedies have led to calls in international fora (G20, 2019; G7, 2019; G20, 2017; Christchurch Call , 2019) to increase efforts to limit the spread of TVEC online in a way that is transparent, accountable and compatible with fundamental rights and freedoms. Industry supporters of the Christchurch Call, for example, have committed to "[i]mplement regular and transparent public reporting, in a way that is measurable and supported by clear methodology, on the quantity and nature of terrorist and violent extremist content being detected and removed" (Christchurch Call , 2019). The 2019 G20 Osaka Leaders' Statement on Preventing Exploitation of the Internet for Terrorism and Violent Extremism welcomed "online platforms' commitment to provide regular and transparent public reporting" (G20, 2019). Greater transparency will improve understanding and assessment of content-sharing services' TVEC-related policies and actions, including content moderation. It will also help to ensure that fundamental rights such as privacy, freedom of expression and due process are not unduly curtailed.

This second benchmarking report explores the degree to which the world's top 50 online content-sharing services' approaches to TVEC online have changed and evolved over the course of one year. As the first benchmarking report did, this report provides an objective snapshot in time. It also informs efforts led by the OECD, in collaboration with member countries, businesses, civil society and academia, to develop a multi-stakeholder, consensus-driven framework and set of metrics for voluntary transparency reporting on TVEC online by content-sharing services. The framework and metrics are intended to lead to a standardised template that any company wishing to report on TVEC can use, and that all OECD members can support.

The key findings of this second benchmarking report are:

- Overall, the degree of transparency and clarity in the top 50 services' TVEC-related policies and procedures has improved appreciably. Six more services began issuing TVEC-specific transparency reports since last year.
- The five services that already issued reports on TVEC now provide additional information, although the nature of that information varies.
- Aside from Microsoft's, the initial reports from the newcomers are relatively basic, but the fact that they have started to report demonstrates meaningful progress and their reports may become more comprehensive over time.
- Although the number of services that publish transparency reports on TVEC is growing, there is still a lack of uniformity in how they define TVEC and related concepts, what and when they report, as well as how they measure or calculate their reported metrics. Therefore, a clear and complete cross-industry perspective on the effects of the

reporting services' efforts to combat TVEC online – including their impact on human rights – cannot currently be obtained. Continued growth in the number of relevant services issuing TVEC transparency reports, as well as greater convergence in the reports' metrics, methodologies, and frequency, would help to bring that perspective into focus.

- Similarly, the number of jurisdictions that have TVEC-related laws and regulations in force or under consideration is growing, but they are not consistent, either. That presents a risk of divergent reporting standards and requirements coming into effect.
- The COVID-19 pandemic and associated lock-down measures have led some Services to increase reliance on automated monitoring systems to detect and remove TVEC.
- Fourteen of the top 50 online content-sharing services surveyed in this report are based in or owned by parent companies based in the People's Republic of China (hereafter 'China'), as compared to thirteen last year. One of them, TikTok, joined the group that issues transparency reports specifically on TVEC since the first benchmarking report was published.
- Because some terrorist and violent extremist groups are moving to small platforms that lack the resources and expertise necessary to moderate TVEC effectively, further research could usefully focus on the services that these groups exploit the most for the purpose of TVEC dissemination, rather than on the global top 50 per se. It could also address their responses to the TVEC that appears on their properties, and potential cooperation and assistance mechanisms through which experienced online platforms could aid these smaller services in their efforts to tackle TVEC online effectively.

Introduction

Whilst the Internet has undeniably brought about significant improvements to our lives, such as expanding cross-border commerce, reducing search and transaction costs, and opening new avenues for communication, it has also posed some new challenges. Amongst them is the use of online content-sharing services to post and disseminate terrorist and violent extremist content ("TVEC").

Terrorist and violent extremist groups have shown a willingness to employ new technologies for recruitment, dissemination of propaganda, communication and mobilisation (United Nations Office on Drugs and Crime, 2012). In addition to using social media to amplify their messages, bad actors can use encrypted messaging apps for intra-group communications and coordination of terrorist attacks, taking advantage of the privacy protection such apps afford to avoid detection (Clifford & Powell, 2019). As the Christchurch and Halle attacks demonstrated, improvements in mobile data infrastructure enabled perpetrators of violent extremist and terrorist acts to share their acts of violence unedited, unredacted and in real time, with audiences accessing them through their smartphones (Ahmed, 2020).

In response, major tech companies like Google, Facebook, Twitter and Microsoft have formed an alliance and taken a number of steps to halt the proliferation of TVEC on their online platforms and prevent the abuse of their service. As seen in Section 3 of the first benchmarking report, several online platforms partnered to form the Global Internet Forum to Counter Terrorism (GIFCT), now an independent body which liaises with governments, civil society, academia and international organisations to combat the spread of TVEC online more efficiently. Among their efforts is a database of known terrorist and violent extremist content hashes – essentially, digital fingerprints consisting of a sequence of letters and numbers – that have already been removed by at least one participating company. The database makes it possible for companies to choose, according to their individual policies, to rapidly prevent the same content from being re-uploaded or to find copies that already exist on any of the other participating services. Also, many online platforms and other Internet-based services have explicitly banned the use of their technologies to support or engage in terrorist and violent extremist activities,[1] taking both proactive and reactive measures to prevent and minimise violations of their terms of service or community guidelines. These measures range from warnings, content removal, and account suspensions to permanent bans from the relevant service.[2]

There is, however, significant variation not only in how platforms moderate TVEC, but in the degree of their transparency about the TVEC that appears on their services and how they address it. There are also transparency and accountability concerns around what is perceived as online platforms' "increasingly aggressive moderation of user-generated content", which could impinge upon individuals' fundamental rights and freedoms such as the right to privacy, free speech and due process (The Santa Clara Principles, n.d.). In response, some technology companies have committed to provide greater transparency in setting their terms of service/community standards, as well as in the manner in which they enforce them, and some governments have pledged to work in tandem with tech companies to devise strategies aimed at preventing the uploading and dissemination of TVEC on social media and similar content-sharing services. These commitments and pledges are set out in the Christchurch Call (2019) and are echoed in a number of international calls for action to eradicate TVEC online in a manner compatible with the protection of individuals' human rights (G20, 2019) (G7, 2019).

In reply to the aforementioned calls for action, the OECD launched a multi-faceted project to develop a framework and set of metrics for voluntary transparency reporting on TVEC, based on an international, multi-stakeholder consensus. Part of this project consists of two "benchmarking" reports, to be issued one year apart. These reports provide snapshots of the TVEC-related policies and procedures of the world's top 50 online platforms and other online content-sharing services (the "Services"), identifying commonalities, developments and trends in the Services' approaches. Emphasis is placed on whether, and if so to what extent, the Services issue transparency reports (TRs) on TVEC. The first report was published in August 2020 (OECD). The present report (the "Report") is the second instalment, focusing on the degree to which the Services' approaches to countering TVEC online have changed and evolved over the course of one year. Like the first instalment, this Report provides an objective and factual snapshot of the Services' current policies and procedures for combatting TVEC. It expresses no opinions on the merits of the policies and procedures, nor does it make any recommendations about them. Rather, this Report provides an evidence base for understanding the Services' approaches to curbing TVEC and determining the extent to which their implementation is transparent and accountable.

Importantly, this Report also informs efforts led by the OECD, in collaboration with member countries, businesses, civil society and academia, to develop a multi-stakeholder, consensus-driven framework and set of metrics for voluntary transparency reporting on TVEC online by content-sharing services. The framework and metrics are intended to become part of a standardised template that all companies wishing to report on TVEC can use, and that all OECD members can support.

Section 1 details this Report's research methodology and scope, explaining how it relates to the first benchmarking report. Section 2 summarises the first benchmarking report's key findings and presents the main changes and developments in the Services' approaches to tackling TVEC online over the past year. Section 3 provides an update on the structure and initiatives of the Global Internet Forum to Counter Terrorism (GIFCT). Section 4 surveys the main developments during the last year in legal and regulatory proposals concerning TVEC in OECD jurisdictions. Annex A is a list of the world's top 50 most popular online content-sharing Services. Annex B contains detailed profiles of those Services, focusing on their TVEC-related policies and procedures. Finally, Annex C contains a glossary of terms that are common in transparency reporting on TVEC.

1. Scope, Methodology and Research Design

The first benchmarking report (OECD, 2020) explored the policies, procedures and practices relevant to TVEC of the world's top 50 Services. Those Services include social media platforms, online communications services, file sharing platforms, and other online Services whose businesses enable the uploading, posting, sharing and/or transfer of digital content and/or facilitate voice, video, messaging or other types of online communications. As explained in Section 1 of the first benchmarking report, the Services in the top 50 list were chosen on the basis of their market penetration or "popularity" under the assumption that TVEC disseminated on popular Services is more likely to reach large audiences. Although the number 50 is necessarily arbitrary, there had to be a cut-off point to set the limits of the research being conducted for these reports. However, it is important to note that inclusion in this top 50 list is not necessarily the same as inclusion in a top 50 list of Services with the highest prevalence of TVEC. This report takes the same approach, one year later, to identify developments in the top 50 most popular Services' approaches to combatting TVEC online over the past year. In particular, the Report examines whether there is more or less clarity in how the Services define TVEC and the procedures they follow to detect and address it, whether the number of Services that publish transparency reports on TVEC has changed, and what metrics those reports include.

As in the first report, given the absence of a common metric that could establish the popularity of all the surveyed Services, this report followed a two-step approach to determine which Services to include in the scope of the research. First, the Services were organised into three categories:

a. social media, video streaming Services and online communications services;
b. cloud-based file sharing Services; and,
c. an "other" category, which includes a content management Service and an online encyclopaedia.

Within each category, the most popular Services were determined based on the following methodology:

- Social media platforms, video streaming Services and online communications Services were identified based on their monthly average users (MAU). The MAU metric is commonly used by industry analysts and investors to determine a service's popularity and growth[3], and constitutes a reliable measure to rank with a fair degree of precision the relative size of Services that thrive on user engagement.
- Cloud-based file sharing Services were identified based on indicative market shares, a metric that is frequently used to determine the relevance of firms in a given industry segment.
- The third category includes a content management system and an online encyclopaedia. The popularity of these two Services cannot be determined relative to the other two groups; however, their undoubted relevance warranted their inclusion. Their importance was determined on the basis of data (indicative market share and monthly pageviews) that reveal their reach and/or usage.

A list of the world's top 50 Services is included in Annex A. Relative to the list in the first report, this year's top 50 list changed little. Other than some fluctuations in the rankings – such as TikTok and Telegram climbing - the only noteworthy changes are the inclusion of the Chinese short video app Kuaishao at number 15 and the exit of the social media platform MySpace.

The review of the Services' approaches to combatting TVEC online consisted of three steps. First, the standardised profile template produced in the first benchmarking report[4] was used to profile each Service. One profile per Service was produced based on each Service's publicly available terms of service (ToS), community guidelines and policies, blogs, service agreements and other official information ("governing documents")[5]. The Services were contacted and given adequate time to provide feedback on the accuracy of their profiles, as well as any relevant additional information.

Secondly, the profiles were updated based on the Services' responses. The final versions of the profiles appear in Annex B.

Thirdly, the main findings of the first benchmarking report were updated based on the newly-compiled information in the Services' profiles. An updated factual and objective overview of the world's top 50 Services' approaches to tackling TVEC online is presented in Section 2 of this Report.

Section 2 focuses on the changes and developments in the Services':

a. policies concerning terrorist/terrorism and violent extremist/violent extremism;
b. detection and removal of TVEC, including policies on enforcing compliance with terms and conditions of service, on removals, on sanctions, and whether there are appeals processes;
c. consequences for user breaches of terms of service/community guidelines and standards;
d. voluntary issuance of transparency reports (TRs) concerning TVEC including their content, methodology and frequency.

2. Updated Commonalities, Developments and Trends in the Services' Approaches to TVEC

Different Descriptions of TVEC and Related Concepts Remain, as Do Diverging Approaches to Identifying 'Terrorist Organisations'.

The first benchmarking report found dissimilar approaches in the Services' content policies and definitions concerning TVEC and related concepts, as well as in their understanding of what amounts to a terrorist group or organisation, the provision of detailed explanations and examples being the exception rather than the rule. Table 1 shows that over the course of the last year, only minor changes were observed.

Table 1. Services' Approaches to Defining TVEC and Related Concepts

Approach	1st benchmarking report	2nd benchmarking report
Services that define terrorism, violent extremism and related concepts with sufficient detail to understand the scope of such terms, providing examples where appropriate	5[6]	6[7]
Services that explicitly ban the use of their technologies to foster terrorist and/or violent extremist aims, using (but not explaining in detail) the terms terrorist/terrorism, violent extremists/violent extremism and similar expressions	19[8]	21[9]
Services that include TVEC within the same reporting category as hate speech and/or violent or graphic content	15[10]	13[11]
Services that use broad and/or general descriptions of prohibited conduct, which descriptions can be interpreted as supersets encompassing TVEC	16[12]	15[13]

Sources: Annex B in (OECD, 2020); Annex B in this Report.

Some of the Services made efforts to clarify what they consider TVEC to be, as well as its unacceptability on their platforms. In particular:

a. Pinterest updated what it deems “dangerous organisations and individuals”.
b. Twitch updated its “Terrorism and Extreme Violence” guidelines, providing greater clarity on how it defines terrorist organisations and how its internal safety teams categorise related content. These clarifications broadened the definition of content that fits in this category (including forms of behaviour in this category that were previously categorised as other types of abuse).
c. Discord has a new prohibition against violent extremism, defined as “content where users advocate or support violence as a means to an ideological end.”
d. Microsoft issued a Digital Safety Content Report (which encompasses Skype and OneDrive), where it clarifies that “both terrorist and violent extremist content is prohibited on Microsoft platforms and services”, and that the Microsoft Services Agreement Code of Conduct prohibits the “posting of terrorist or violent extremist content”.

The first benchmarking report concluded that the Services had different approaches to identifying and defining a terrorist organisation. This finding remains valid one year later. Some Services like Facebook and Instagram use their own definitions of terrorist organisations, distinguishing them from hate organisations, mass and multiple murderers, human trafficking groups and criminal organisations[14]. Other Services like those provided by Microsoft, YouTube, Wordpress.com and Quora rely on United States government or United Nations lists of terrorist organisations[15]. VK follows the legal definition of terrorist content provided in the countries where it has a presence[16]. The majority of the Services, however, provide no information in this regard[17].

Transparency Reports Expressly Addressing TVEC Are Still Uncommon among the Top 50, but There Are Several Possible Explanations

One of the main findings of the first benchmarking report was that of the 23 Services that issued TRs of any kind, only five (Facebook, YouTube, Instagram, Twitter and Automattic) issued reports specifically about TVEC. In the last year, Skype, OneDrive, Twitch, TikTok, Reddit and Discord joined the group of Services that provide information on TVEC removals in their TRs.

TikTok stands out for being the first Chinese-owned Service to publish TRs of any kind, and now for publishing TRs about TVEC specifically. This effort was accompanied by the release of new Community Guidelines and the launch of a “Transparency Center” (Perez, TikTok to open a 'Transparency Center' where outside experts can examine its content moderation practices, 2020).

It is important to note that malicious actors are not using all of the global top 50 online content-sharing Services to spread TVEC, which could explain why not all of the Services issue TVEC-specific TRs. For example, Pinterest, Medium and Meetup do not seem to be places where much, if any, TVEC is surfacing, at least for the time being. Moreover, TVEC is not evenly or even proportionately disseminated amongst the online content-sharing Services where it does appear. Thus, whilst some Services like 4chan, Telegram and YouTube may detect substantial volumes of TVEC, the sizes of their user bases vary significantly. Meanwhile, other Services such as Wikipedia and LinkedIn are highly popular but seem to be rarely used for TVEC-related purposes.

However, as seen in Section 11 of the Profiles listed in Annex B, TVEC appeared in at least 27 Services at some point in time, a number which is significantly larger than the 11 Services that have issued TVEC-specific TRs to date. Then again, there are factors other than being TVEC-free that may explain why some Services do not issue TVEC-specific TRs. For example, 13 Services are Chinese platforms, which are impeded from issuing TRs due to tensions between local regulatory requirements and business considerations (see below paragraphs 43-47). Also, end-to-end encrypted Services like iMessage/Facetime, Telegram and WhatsApp cannot see

the content of their users' communications, for which reason they are naturally prevented from issuing sufficiently informative and detailed TVEC-specific TRs[18].

The absence or relative scarcity of TVEC, regulatory constraints and technical considerations may explain a service's lack of motivation or inability to issue TRs on TVEC. Nevertheless, TVEC policy-making efforts could benefit if there were more clarity in this regard. For example, if companies stated that they do not issue TRs on TVEC because no TVEC appears on their services, it would be easier to narrow down the universe of Services which have a significant role in the dissemination of TVEC. It is also the case that some of the services that do have such a role are not necessarily those with the largest user bases, that is, those which are the main focus of the two benchmarking reports.

Indeed, experts recommend that any response to the problem of TVEC online not be limited to just a few large platforms. Rather, they urge that the problem must be seen as a whole – i.e. considering how the response from large platforms leads to changes in their usage patterns for TVEC dissemination, such as mass migration to more obscure platforms, services and apps, including those hosted on the Dark Web (Tech Against Terrorism, 2019). Since research has shown that some terrorist and violent extremist groups are moving to small-sized platforms that lack resources and expertise adequate to police TVEC effectively (Tech Against Terrorism, 2019), further research could usefully focus on the services that these groups exploit the most for the purpose of TVEC dissemination. It could also address their responses to the TVEC that appears on their properties, and potential cooperation and assistance mechanisms through which experienced online platforms could aid these smaller services in their efforts to tackle TVEC online effectively.

Differences between TVEC Transparency Reports Remain, but They Include More Information Now

The first benchmarking report showed that the definitions used and the kinds of information included in the five TVEC TRs then issued were largely different from one another. This remains the case one year later. However, it is possible to discern a general trend among the five Services that were publishing TVEC TRs last year towards providing additional information.

Twitter, for example, in addition to reporting the accounts actioned and accounts suspended for violating the Twitter Rules, including the policies against terrorism and violent extremism, now reports the "content removed" metric, i.e. the number of unique pieces of content (such as Tweets or an account's profile image, banner, or bio) that Twitter required account owners to remove for violating the Twitter Rules. Twitter also includes trends in the reported data, some of which concern TVEC. For example, in its last report, Twitter observed that there was a 9% decrease in the number of accounts actioned for violations of its Terrorism and Violent Extremism Policy as compared to the last reporting period.

Similarly, in addition to the metrics in its earlier benchmarking reports, YouTube now includes the total number of appeals received per quarter for videos removed due to a community violation and the total number of videos that YouTube reinstated per quarter due to an appeal after removal for a community guidelines violation. It also discloses the percentage of all video removals that occurred before any users viewed the removed video and the percentage of all removals that occurred after the video was viewed (known as the "violative view rate").

Facebook continues reporting the same metrics as last year: 1) how prevalent terrorist propaganda violations on Facebook were; 2) how much content Facebook took action on; 3) the percentage of the violating content Facebook actioned before users reported it; 4) the number of appeals against the decisions to take an action on specific content; and, 5) the amount of content Facebook restored after removing it. However, Instagram, which last year was reporting the first three metrics, now reports all of them, as well. Furthermore, both Facebook and Instagram are also reporting recent trends regarding content actioned for organised hate and terrorism. For example, Facebook's last TR notes that content actioned for organised hate decreased from 4.7 million pieces of content in Q1 2020 to 4 million in Q2 2020,

and content actioned for terrorism increased from 6.3 million pieces of content in Q1 2020 to 8.7 million in Q2 2020. Moreover, Facebook updated its document titled "Understanding the Community Standards Enforcement Report", now providing highly detailed explanations on how both Facebook and Instagram moderate content and calculate the metrics they report.

Automattic's (Wordpress.com's parent company) reported metrics remained the same. However, in the Summary section of its TR, Automattic now reports the number of sites/content specified in the Internet Referral Unit notices for the period 1 January 2018 to 30 June 2020.

Microsoft began publishing a Digital Safety Content Report, which encompasses Microsoft consumer products and services including (but not limited to) OneDrive, Skype –services that are profiled in the two benchmarking reports – as well as other products such as Bing, Xbox and Outlook. In its Report, Microsoft reports TVEC-specific metrics, including the amount of TVEC actioned, the number of accounts suspended due to TVEC, the percentage of TVEC actioned that Microsoft detected, the percentage of TVEC actioned reported by users or third parties, and the percentage of accounts suspended for TVEC that were reinstated upon appeal.

Twitch published its first TR in February 2021, covering the first and second halves of 2020 (Twitch, 2020). The TR explains Twitch's efforts and methods to enforce its Community Guidelines and provides information on the extent to which TVEC appears on its platform. Twitch explains that because it is a live-streaming service, the vast majority of its content is ephemeral. Live content is flagged by either machine detection or user reports to Twitch's team of content moderators (i.e. paid staff), who then issue "enforcements" (typically a warning or timed channel suspension) for verified violations. If there happens to be recorded content that accompanies a violation, that content is removed. But most enforcements do not require content removal because, apart from the report, there is no longer a record of the violation. For this reason, Twitch does not focus on "content removal" as the primary means of enforcing adherence to its Community Guidelines. Rather, the number of "enforcements" is a better measure of its Community Guidelines enforcement efforts. Accordingly, Twitch reports the number of enforcements it issued for violations of different categories, one of which is "Terrorism, Terrorist Propaganda and Recruitment".

TikTok, Reddit and Discord also began to publish TRs on TVEC since last year's report was written. The scope of these initial reports, however, is comparatively modest.

a. TikTok reports the percentage of videos removed for violation of its policies on hate speech, integrity and authenticity and dangerous individuals and organisations;
b. Reddit reports the number of pieces of designated foreign terrorist organisation content (designated foreign terrorist organisation is not defined) it removed during the reporting period; and
c. Discord reports the number of violent extremism server deletions by month, proactively detected with automated tools.

Overall, the developments noted above, especially the information provided by Twitch in its first TR, help to clarify the reporting Services' efforts to fight TVEC. Also, Twitch's updated guidelines on "Terrorism and Extreme Violence" and Facebook's updated "Understanding the Community Standards Enforcement Report" document signal that they are taking TVEC seriously, and reflect a commitment to keep users informed on how they address it. In addition, the fact that six more Services now provide information on TVEC suggests that the significance of the fight against it, as well as the need for transparency about it, are becoming more widely acknowledged.

The numbers reported by Twitch, Microsoft, TikTok, Reddit and Discord, in addition to Twitter's new "content removed" metric, shed additional light on the prevalence of TVEC and, to a limited extent, signal some convergence in transparency reporting on TVEC. However, Microsoft reports TVEC metrics in aggregate for all Microsoft consumer services and products (which are not listed exhaustively), and not on a per-product and per-service basis. This approach provides helpful information at the corporate level, but does not allow a perspective on the distribution of TVEC amongst Microsoft's products and services. Moreover, as seen above, the information provided by TikTok, Reddit and Discord is quite narrow, and the variance amongst the other

five Services' reporting approaches remains high. As a result, there is still very limited room for comparison and analysis across reports. All the observations made in this section of the first benchmarking report remain pertinent and valid[19].

Because the number of Services that publish TRs on TVEC remains low, and there is still a lack of uniformity in what they report and how they calculate their metrics, a clear and complete cross-industry perspective on the nature and efficacy of the Services' measures to combat TVEC online – as well as on the human rights impact those measures have – cannot currently be obtained. A higher number of Services issuing TVEC TRs, as well as greater convergence in the metrics they report and their calculation methodologies, would enable better assessments of the overall nature and impact of the Services' TVEC policies and moderation practices.

Staff Member Moderators, User-Moderators and Automated Tools: Heavier Reliance on Automation in the COVID-19 Era

The first benchmarking report showed that the Services had different approaches to moderating TVEC, relying on staff member moderators, user-moderators, automated tools, or a combination of them. Table 2 shows that there were no significant changes in these approaches over the course of the last year.

Table 2. Services' Content Moderation Methods

Method	1st benchmarking report	2nd benchmarking report
Services that rely on staff member moderators	40[20]	40[21]
Services that rely on user-moderators	10[22]	10[23]
Services that rely on automated tools	At least[24] 21[25]	At least[26] 23[27]

Note: These methods are not mutually exclusive. That is to say, they can be used in combination.

Sources: Annex B in (OECD, 2020); Annex B in this Report.

It is noteworthy, however, that as a result of the COVID-19 pandemic and lock-down measures, some Services, such as Facebook, YouTube and Twitter, increased their reliance on automated monitoring systems to flag and remove problematic content, including TVEC. These systems cannot make the nuanced judgments that are sometimes necessary to determine whether specific content amounts to TVEC (Duarte, Llanso, & Loup, 2017). Accordingly, to reduce the likelihood of missing such content, these systems tend to be programmed to err on the side of caution. That, in turn, raises the risk of false positives, as recognised by YouTube, which noted that as a consequence of greater reliance on automated moderation systems, "users and creators may see increased video removals, including some videos that may not violate [its] policies" (YouTube, 2020). Some commentators have observed, for example, that Arabic-language content is disproportionately flagged as terrorist content (Stokel-Walker, 2020). Therefore, there is a risk that content moderation without human oversight may exaggerate the volume of TVEC that is showing up. On the other hand, there is also a possibility that some TVEC is going undetected due to algorithmic flaws or effective concealment. At any rate, the

magnitude of false positives and false negatives is unclear, as is whether such errors have an appreciable influence on the relevant Services' reported metrics.

The limitations of automated tools to moderate TVEC became apparent during the Christchurch attack. According to Facebook, the video of the attack in Christchurch did not prompt its automatic detection systems because it did not have enough content depicting first-person footage of violent events to effectively train its machine learning technology at that time. Accordingly, Facebook started working with government and law enforcement officials in the United States and the United Kingdom to obtain camera footage from their firearms training programs – providing a valuable source of data to train its systems. With this initiative, Facebook aims to improve its detection of real-world, first-person footage of violent events and avoid incorrectly flagging other types of footage such as fictional content from movies or video games (Facebook, 2019).

Notification, Enforcement and Appeal Mechanisms Remain Varied

The notification of enforcement decisions, as well as the possibility to appeal them, are important measures for safeguarding due process. The first benchmarking report showed that the Services' approaches to notifications and appeals were not uniform. Table 3 illustrates that this is still the case one year later.

Table 3. Services' Approaches to Notifications and Appeals

Approach	1st benchmarking report	2nd benchmarking report
Services that have mechanisms for notifying users in case of potential violations of their ToS and other governing documents	21[28]	23[29]
Services that have appeal processes in place in respect of content moderation decisions and other measures applied under their governing documents	23[30]	27[31]

Sources: Annex B in (OECD, 2020); Annex B in this Report.

The remaining Services either have no appeal processes or do not provide public information in this regard.

The first report also showed that with respect to 22 Services it was difficult to obtain a clear understanding of whether they reviewed content proactively and/or reactively to determine compliance with their ToS and policies[32]. After one year, this number declined to 17[33].

Disclosure by Chinese Platforms

Lastly, the first report noted that the Chinese Services generally provided limited information with respect to their content moderation practices and processes for enforcing their ToS and policies. With the exception of TikTok, none of them issued TRs of any kind.

As noted above, TikTok began issuing TVEC TRs. There were no changes in the remaining Chinese Services' disclosure approaches over the course of the last year[34].

The first benchmarking report explained that the Chinese Services' limited disclosures regarding content moderation and monitoring may be the result of striking a balance between the Services' obligation to comply with local laws and regulations (under which they are bound to monitor and censor content in close co-operation with the Chinese government) and their need to keep their services attractive. Admitting content monitoring and moderation to comply with local rules could make their services unappealing due to lack of privacy and censorship of speech. In spite of the narrow disclosures, the fact that Internet-based Chinese Services monitor and censor content in collaboration with the government is well documented. For example, research has shown that platforms like WeChat implement different keyword-based and scanning tools to monitor content and improve China's surveillance mechanisms (Ruan, Knockel, Ng, & Crete-Nishihata, 2016). Political activists also report having been followed based on what they have said on WeChat, and chat records have turned up as evidence in court (Zhong, 2018). Chinese social media platforms and apps play a paramount role in the implementation of China's "social credit system", largely deemed a mass surveillance and governmental control system in Western societies (Lix Xan Wong & Shields Dobson, 2019). This co-operation is enabled by many laws passed in furtherance of state security, public security, censorship and taxation that have granted the Chinese government extensive powers of access to private-sector data generated online by businesses operated in China (Wang, 2017).

Concerns about the possibility that Chinese Services are components of the Chinese government's surveillance system may hinder any ambitions for international expansion that those Services have. The first report noted that some Chinese Services have made significant efforts to dispel such concerns outside China. In particular, TikTok has undertaken a number of initiatives to increase transparency in its content moderation and removal practices (Perez, TikTok to open a 'Transparency Center' where outside experts can examine its content moderation practices, 2020), ensuring it is not conflated with its Chinese version, Douyin, and claiming that no user data is ever sent to China (Cuthbertson, 2019). WeChat likewise implemented a "one app, two systems" censorship model, under which only WeChat users with accounts registered to mainland Chinese phone numbers are monitored and censored (Ruan, Knockel, Ng, & Crete-Nishihata, 2016).

However, some sources suggest that TikTok's and WeChat's assurances are misleading. Recent research has shown that WeChat monitors non-China-registered accounts and uses messages from those accounts to train censorship algorithms to be used against China-registered accounts (Kenyon, 2020). Similarly, a white paper by the cybersecurity firm Penetrum found that over one-third of the IP addresses the TikTok Android Package Kit connects to are based in China, concluding that "TikTok does an excessive amount of tracking on its users and the data collected is partially if not fully stored on Chinese servers with the ISP Alibaba" (Penetrum Security). It has been reported that searches on TikTok revealed significantly fewer videos of the Hong Kong (China) protests than expected – thus suggesting censorship is taking place (Harwell & Romm, 2019). Furthermore, content moderation guidelines advancing Chinese foreign policy through the TikTok app were leaked last year (Hern, 2019). On account of its use of Chinese infrastructure and its parent company's close ties to the Chinese Communist Party, ex-European Organisation for Nuclear Research (CERN) security engineers recently warned that TikTok is a "perfect tool for massive surveillance and data collection by the Chinese government" (Kock, 2020). Such concerns only enhance the desirability of thorough transparency reporting on these Services' content moderation practices.

3. Update on the GIFCT

The Global Internet Forum to Counter Terrorism (GIFCT) was founded by Facebook, Microsoft, Twitter and YouTube in 2017 to curb the spread of TVEC on digital platforms. Please see Section 3 of the first benchmarking report (OECD, 2020) for an overview of the GIFCT's goals and initiatives.

The Global Internet Forum to Counter Terrorism (GIFCT) has been slowly but steadily expanding since its foundation by Facebook, Microsoft, Twitter and YouTube in 2017. After Amazon, Dropbox, LinkedIn, Pinterest, WhatsApp and Instagram became members in 2017-2019, Mega.nz, Mailchimp and Discord are the latest entities to join (GIFCT, 2021).

Under its new structure agreed in 2019, the GIFCT is governed by an Operating Board, which works closely with a broad Multi-Stakeholder Forum and an Independent Advisory Committee. The Operating Board hires the Executive Director, provides the initial operational budget, and ensures overall GIFCT operations align with its mission. The Operating Board is composed of:

- GIFCT's founding members - Facebook, Microsoft, Twitter, and YouTube
- At least one rotating company from the broader membership cadre
- New companies that meet leadership criteria
- The rotating chair of the Independent Advisory Committee, who participates as a non-voting member

The Operating Board chair rotates annually. Microsoft is the Operating Board chair for 2020 (GIFCT, 2020).

The Multi-Stakeholder Forum includes a wide range of companies, civil society members and governments committed to upholding and respecting human rights and preventing terrorists from exploiting digital platforms. The Forum serves as the primary vehicle for information-sharing and ideas-exchange to help guide GIFCT activities and engagement (GIFCT, n.d.).

On 16 June 2020, the GIFCT announced full membership of the inaugural Independent Advisory Committee (IAC). The 21 members include representatives from seven governments, two international organisations, and 12 civil society organisations (CSO), spanning a range of expertise.

The GIFCT Hash Sharing Consortium, which shares "hashes" or "digital fingerprints" of known terrorist images and videos, maintains a database with about 300 000 unique hashes. These consist of approximately 250 000 visually distinct images and approximately 50 000 visually distinct videos. The Consortium is composed of 13 companies that have access to the shared industry database: Microsoft, Facebook, Twitter, YouTube, Ask.fm, Cloudinary, Instagram, JustPaste.it, LinkedIn, Verizon Media, Reddit, Snap and Yellow (GIFCT, 2020). Hashes are labelled following the taxonomy below:

- Imminent Credible Threat (ICT): A public posting of a specific, imminent, credible threat of violence toward non-combatants and/or civilian infrastructure.
- Graphic Violence Against Defenceless People: The murder, execution, rape, torture, or infliction of serious bodily harm on defenceless people (prisoner exploitation, obvious non-combatants being targeted).
- Glorification of Terrorist Acts (GTA): Content that glorifies, praises, condones or celebrates attacks after the fact.

- Recruitment and Instruction (R&I): Materials that seek to recruit followers, give guidance or instruct them operationally.
- New Zealand Perpetrator Content: Due to the virality and cross-platform spread of the Christchurch attacker's manifesto and attack video, and because New Zealand authorities deemed all manifesto and attack video content illegal, the GIFCT created a crisis bank in the hash database to help mitigate the spread of this content.
- Halle, Germany, Perpetrator Content: On 9 October 2019 the GIFCT activated its new Content Incident Protocol (CIP) for the first time after the protocol's development following the terrorist attack in Christchurch, New Zealand the previous March. The CIP was declared following the tragic shooting in Halle, Germany and the circulation of the perpetrator's attack video on multiple digital platforms.
- Glendale, Arizona, U.S., Perpetrator Content: On 20 May 2020 the GIFCT activated its Content Incident Protocol following the shooting in Glendale, AZ, adding hashes of visually distinct videos depicting the attacker's content during the shooting (GIFCT, 2020).

The Hash Sharing Consortium launched a new feature in 2019 to better allow consortium members to express disagreement with hashes shared within the database. If a company believes that a hash in the database was added erroneously or has been mislabelled, they can express that disagreement in two ways. First, a company can add a label indicating agreement that the hash is terrorist content, but that they believe it was labelled incorrectly via the taxonomy. Second, a company can add a label to a hash indicating that they do not feel the content is explicitly terrorist content (disputed content). These labels are visible to all companies within the Hash Sharing Consortium so that third companies can make their own decision on how best to use the hashes within various taxonomy buckets, based on their own processes and review systems (GIFCT, 2020).

The GIFCT has stated that its Content Incident Protocol (CIP) is a process by which GIFCT member companies become aware of, quickly assess, and act on potential content circulating online resulting from a real-world terrorism or violent extremist event. Since the Christchurch tragedy, GIFCT member companies have developed, refined and tested the protocol. The CIP assessment process was initiated more than 100 separate times between March 2019 and November 2020 (GIFCT, n.d.).

No individual or organisation can activate a content incident. Rather, the protocol is based on the existence of content online relating to a real-world terrorism or violent extremism event – like Christchurch, Halle or Glendale – and potential distribution of that content, featuring a livestream of murder or attempted murder produced by the attack's perpetrator or an accomplice. The CIP is a multi-step process, including a decision to initiate the CIP, communication of that decision, a review of content assets, and other steps, to inform GIFCT member companies and affected governments about content from the real-world event that may be manifesting online. A CIP ends with an official "conclusion" determined by GIFCT founding member companies once the volume of content has noticeably decreased across GIFCT member platforms (GIFCT, n.d.).

GIFCT has also initiated a set of working groups, one of which focuses on transparency. The transparency group has met monthly since July 2020, includes representatives from business, government, international organisations, academia, and civil society, and is undertaking a programme of work to improve the understanding and utility of transparency.

Lastly, during the past year, the GIFCT adapted its URL-sharing programme through a 12-month URL sharing pilot with SITE Intelligence, a firm that provides subscription-based monitoring and analysis regarding terrorist content and other online harms. The pilot project gave some of GIFCT's newer members access to SITE's SourceFeed, providing access to a dashboard assisting with extra context around a given URL, including organisational affiliation of the terrorist content and translation of content into English and further context support. Through this programme, the GIFCT has now shared nearly 24,000 URLs since its launch (GIFCT, 2020).

4. TVEC-related Laws and Regulations that Are in Force or under Consideration

Social media and other online communications services have been identified[35] as integral tools to terrorist and violent extremist groups' recruitment, engagement, and coordination efforts. Moreover, information shared on these platforms is perceived by individuals who are at risk of becoming extremists as more reliable than news media because the content is not framed by the perceived biases of media outlets.[36]

Because terrorist and violent extremist groups misuse online services to disseminate propaganda and recruitment material, technology companies have faced increased pressure from governments and institutions around the world to ramp up efforts to combat the groups' operations. Concerned that, to date, industry efforts to counter TVEC have been inadequate, some governments have begun to propose and enact laws and regulations, and to implement other initiatives, to curb the online propagation of TVEC. Incidentally, just as the Services' policies and approaches to transparency reporting on TVEC vary, so do the legislative and regulatory responses to TVEC. Consequently, there is a lack of coordination on both sides of the coin. This Section provides an overview of TVEC-related laws and regulations that have been enacted or that are currently under consideration.

Australia

In the aftermath of the Christchurch terrorist attacks, the Australian Parliament responded by passing the *Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019* (Act), which came into force on 6 April 2019 (Australian Government, 2019). The Act adds new offences to the Criminal Code concerning online abhorrent violent content.

Abhorrent violent material is audio, visual, or audio-visual content that records or streams abhorrent violent conduct, produced by the perpetrator(s) of that conduct (or an accomplice) that a reasonable person would consider offensive in the circumstances. Abhorrent violent conduct is defined to mean murder or attempted murder, a terrorist act, torture, rape or kidnapping. There is no requirement that the person needs to be convicted of an offence in order for their conduct to constitute abhorrent violent conduct. For the purposes of the Act, it is immaterial whether or not the abhorrent violent material has been altered (for example, through the superimposition of other material). However, if the material is altered (through appropriate editing) to such an extent that it no longer meets the criteria of abhorrent violent material, it will not be captured by the legislation.

Under the Act, it is an offence for an Internet service provider, content service or hosting service to fail to refer to the Australian Federal Police (AFP) "within a reasonable time" abhorrent violent material that the provider is aware could be accessed through or on their service, where the underlying conduct occurred or is occurring in Australia. The term "reasonable time" is not defined in the Act. However, the Explanatory Memorandum states that this will ultimately be a question for the trier of fact (for example, a jury) and will depend on factors such as the volume of the material (for example, how frequently it was posted and re-posted) and the capacity and resources of the service provider (that is, its technical removal capabilities).

In addition, under the Act it is an offence for a content or hosting service provider to fail to expeditiously remove from their content or hosting service abhorrent violent material that is reasonably capable of being accessed in Australia (regardless of where the service itself is

located). The question of whether or not specific content has been "expeditiously removed" is, again, a matter for the trier of fact and will depend on factors such as the type and volume of the material and capabilities and resources of the service provider.

The Act also empowers the eSafety Commissioner to issue notices to content or hosting service providers to notify them that their services could be used at the time of issuing the notice to access abhorrent violent material. This does not create criminal liability. In any subsequent prosecution, the notice will create a presumption that a service provider was reckless about their service being used to access abhorrent violent material. The prosecution is still required to prove the elements of the offence to criminal standard of proof. The presumption can be rebutted by the provider pointing to evidence to the contrary position.

The eSafety Commissioner may employ a direction power under subsection 581(2A) of the *Telecommunications Act 1997* to give written directions to service providers in connection with any of the eSafety Commissioner's powers or functions. In July 2019, the Minister for Communications, Cyber Safety and the Arts conferred a new function upon the eSafety Commissioner via legislative instrument to promote online safety for Australians by protecting them from access or exposure to material that promotes, incites, or instructs in, terrorist acts or violent crimes.

The eSafety Commissioner used the direction power in September 2019 to formalise action already taken by Internet service providers to block websites known to be providing access to footage of the Christchurch attacks and/or the perpetrator's manifesto. The blocking direction was a temporary measure issued following an industry consultation process, which included providing website administrators with an opportunity to remove the content voluntarily. The direction expired in March 2020.

Moreover, the Australian Taskforce to Combat Terrorist and Extreme Violent Material Online (the Taskforce) was established in March 2019, the objective of which is to provide advice to Government on practical, tangible and effective measures and commitments to combat the upload and dissemination of terrorist and extreme violent material (Department of the Prime Minister and Cabinet, 2019). Fulfilling its remit, the Taskforce issued a report on 30 June 2019, identifying actions and recommendations that fall into one of five streams: prevention; detection and removal; transparency; deterrence; and capacity building. Some of such actions and recommendations include:

a. Digital platforms must continue to develop and report to the Australian Government on the ongoing development of technical solutions that seek to prevent terrorist and extreme violent material from being uploaded onto their services,
    - Industry representatives reported to the Australian Government in September 2019 with implementation reports outlining the actions they intended to take to implement the Taskforce recommendations. The next round of annual reporting from industry representatives were due to the Australian Government in November 2020
b. Digital platforms must work with other members of the GIFCT to strengthen the hash-sharing database and the URL-sharing consortium, with an aim to align, to the extent possible, with the categories of violent content prohibited by platforms under their respective community standards and terms of service, such as graphic violence, violent content or gore.
c. Digital platforms must have in place clear, efficient appeals mechanisms that provide users with the ability to challenge moderation decisions regarding terrorist and extreme violent material.
d. Overseen and managed by the Australia-New Zealand Counter-Terrorism Committee, digital platforms and relevant Australian Government agencies should convene a 'testing event' in 2019-20, simulating a scenario which will allow all parties to gauge whether industry tools, and Government processes, are working as intended, particularly as they mature in response to technology and increased investment in content moderation.

- The testing event was held on 1 October 2020 and resulted in the finalisation of Australia's *Online Content Incident Arrangement*.

e. The eSafety Commissioner, in consultation with the Communications Alliance, should develop a protocol to govern the interim use of the Commissioner's power to direct Internet service providers to block websites hosting offending content in the circumstances of an online crisis event
   - The protocol was finalised in December 2019

f. The Australian Government should pursue legislative amendments to establish a content blocking framework for terrorist and extreme violent material online in crisis events.
   - In December 2019, the Australian Government announced a proposal for a new Online Safety Act, which would include a new content blocking measure. The eSafety Commissioner would be granted the power to direct ISPs to block domains containing terrorist or extreme violent material, for time limited periods, during an online crisis event. This power would be more targeted than the eSafety Commissioner's existing blocking power and provide ISPs with civil immunity when acting in accordance with a blocking direction. The Australian Government is considering feedback from public consultation as the online safety reform package advances.[37]

g. Digital platforms should publish reports (at least half yearly) outlining their efforts to detect and remove terrorist and extreme violent material on their services. These reports are intended to demonstrate the nature and extent of actions being taken by platforms, and could include:
   - the number of items flagged by users for potential violations of policies against the promotion of terrorism or extreme violent content;
   - the total number of items removed by the digital platform
   - the number and entity type (e.g. video, channel) of items of terrorist content and extreme violent content removed by the platform;
   - examples of content flagged for promotion of terrorism or extreme violence that did and did not violate the platform's guidelines;
   - the number of items of terrorist content and extreme violent content that were flagged or identified by the platforms' systems;
   - the total number of items of terrorist content and extreme violent content that were subject to moderation, broken down by those that were flagged by users, systems, other sources, and the total volume of content removed; and
   - the average time taken to review and action flagged items of terrorist content and extreme violent content, or the number of times flagged terrorist content or extreme violent content was viewed by users before action was taken.
   - the implementation of appropriate checks on live-streaming aimed at reducing the risk of users disseminating terrorist and extreme violent material online (Department of the Prime Minister and Cabinet, 2019).[38]

Canada

Canada's current approach to prosecuting TVEC is based on Canada's *Criminal Code*. Canada has a number of criminal offences targeting harmful online behaviours. These include prohibitions on hate propaganda offences of advocating or promoting genocide, incitement of hatred in a public place likely to lead to a breach of the peace, wilful promotion of hatred and terrorism offences. The *Criminal Code* also provides powers for courts to order the removal of certain content from Internet services hosted in Canada. These "take-down" procedures exist in relation to the abovementioned prohibitions.

Canada is currently updating its approach to TVEC on social media platforms. In 2019, the Minister of Canadian Heritage was tasked to "create new regulations for social media platforms, starting with a requirement that all platforms remove illegal content, including hate speech, within 24 hours or face significant penalties. This should include other online harms such as radicalization, incitement to violence, exploitation of children, or creation or distribution of terrorist propaganda". (Government of Canada, 2021)

Relatedly, Heritage Minister Steven Guilbeault is reportedly introducing legislation to create a new government regulator with the power to monitor social media platforms and levy fines on social media companies that allow content like hate speech to remain on their platforms (Thompson, 2021). The government is also set to introduce a 24-hour takedown notice, which would give the regulator power to compel platforms to remove material the regulator deems illegal or hateful, or that otherwise fosters radicalisation, incites violence or promotes terrorist propaganda (Patriquin, 2021).

Development is underway on these new regulations, with an aim to introduce new rules in 2021 for social media platforms. Canada's approach is taking into account current developments and existing regulatory regimes already in place around the world, with a goal to safeguard the safety and wellbeing of Canadians online while preserving freedom of expression.

European Union

On 30 September 2020, the European Commission adopted a Report assessing the measures that Member States have taken to comply with the EU Directive on combating terrorism (European Commission, 2020), including on Article 21, which requires Member States to ensure the prompt removal or blocking of online content hosted in their territory constituting a public provocation to commit a terrorist offence. These measures must be transparent and provide adequate safeguards (including judicial redress) to ensure that they are limited, proportionate and that users are informed of the reason for those measures. Overall, the transposition of this article is uneven across Member States, as some Member States did not transpose Article 21 fully into their national law.

Following a proposal by the European Commission in September 2018, the European Parliament and the Council of the European Union agreed on the "Regulation to address the dissemination of terrorist content online' in December 2020. The Regulation was adopted during the plenary of the European Parliament on 28 April 2021. The obligations under the Regulation are for hosting service providers established in the European Union to address the misuse of their platforms by terrorists. National competent authorities will be able to send orders directly to the companies to remove content within one hour of receiving a removal order. Member States can also require that companies take proactive measures where existing ones are not sufficient to effectively mitigate the risks of terrorist content being disseminated on their services. Hosting service providers will be free to choose the measures they consider most appropriate taking into account their size, capabilities and available resources.

The definition of terrorist content online is in line with the definition of terrorist offences set out in the Terrorism Directive, covering the most harmful content, including material inciting or advocating terrorist offences, such as the glorification of terrorist acts, soliciting a person or a group of persons to participate in the activities of a terrorist group, and providing instructions on how to conduct attacks, including instructions on the making of explosives. Material disseminated for educational, journalistic, artistic or research purposes or for awareness-raising purposes against terrorist activity is protected under the proposed Regulation.

Next to obligations to remove illegal content, the Regulation also includes multiple safeguards to strengthen accountability and transparency about measures taken to remove terrorist content, and against erroneous removals of legitimate speech online. Article 7 of the Regulation introduces transparency obligations for hosting service providers. In particular, they are bound

to set out in their terms and conditions their policy for addressing the dissemination of terrorist content. In addition, they must issue annual transparency reports, including information about the measures taken to identify and remove terrorist content, the use of automated tools, the numbers of content removed or reinstated, and the numbers of complaints and review procedures and their outcomes.

The Commission adopted in December 2020 a proposal for the Digital Services Act (DSA)[39], which aims to clarify the responsibilities and strengthen the accountability of services that intermediate content. The Digital Services Act significantly improves the mechanisms for the removal of illegal content and for the effective protection of users' fundamental rights online, including freedom of speech. It also creates stronger public oversight over online platforms, in particular platforms that reach more than 10% of the EU's population. The proposed measures include:

- measures to counter illegal goods, services or content online, such as a mechanism for users to flag such content and for platforms to cooperate with "trusted flaggers";
- effective safeguards for users, including the possibility to challenge platforms' content moderation decisions, even where they are based on platforms' own terms and conditions;
- transparency measures for online platforms for all content moderation decisions as well as transparency of recommender algorithms of "very large online platforms";
- obligations for very large online platforms to prevent the misuse of their systems for the dissemination of illegal content or intentional manipulation of their service by taking risk-mitigation measures;
- independent audits of the risk management systems as well as access for researchers to key data of the largest platforms, in order to understand how online risks evolve;
- transparency around advertisements towards users as well as an obligation on very large online platforms to maintain ad archives with public access;
- an oversight structure to address the complexity of the online space: EU countries will have the primary role, supported by a new European Board for Digital Services - for very large online platforms, the Commission will be in charge of enhanced supervision and enforcement.

The draft law is currently going through the EU legislative process and is being scrutinised by the Council of the European Union and the European Parliament.

France

On 18 June 2020, the French Constitutional Council ruled on the conformity of a law adopted by the Parliament (the "Avia law") to control the spread of hate speech online. The Council struck down the law's flagship provisions, which would have reduced the time within which platform providers were required to respond to reported hate speech content to (i) 24 hours for hate speech content reported by users and (ii) 1 hour for specific harmful content reported by French authorities as child pornography or terrorist propaganda, subject to criminal sanctions.

The Constitutional Council deemed these provisions unconstitutional on the grounds that they undermine freedom of speech without being "appropriate, necessary and proportionate" to their purpose. The Constitutional Council also found the deadlines too short, the lack of intervention by a French judge problematic, and the risk of "over-censorship" or "over-blocking" troubling. As a result, the Council significantly reduced the scope of the Avia law. The surviving provisions result in:

a. an increase of the criminal fine from EUR 75 000 to EUR 250 000 (to be multiplied by five for legal entities) in case of non-compliance with the following obligations (that already existed under French law) for online platform providers:

i. the obligation to hold and retain data allowing the identification of anyone who creates content through their services (for potential transmission to the French authorities);

ii. the obligation to provide an accessible tool to notify the harmful content promoting crimes against humanity, provoking and promoting acts of terrorism, promoting hatred against persons on grounds of their race, sex, sexual orientation or identity or disability, child pornography, violence, and violating human dignity;

iii. the obligation to promptly inform the competent authorities of any of the above harmful content reported to them and that come from users of their services;

iv. the obligation to make available to the public the means implemented to tackle the content;

b. the creation of a specialised digital prosecutor's office regarding certain types of criminally reprehensible harmful content; and

c. the creation of an "online hate speech observatory" linked to the French broadcasting regulatory authority.

After the Constitutional Council ruling on the Avia law, a bill on "upholding Republican principles" was introduced on 15 January 2021, seeking to subject digital platforms to obligations concerning the moderation of illegal online hate speech. The most controversial provision of the Avia law has been dropped, that is, the injunction for social networks to remove clearly illegal hate content within 24 hours. According to the bill, platforms will have to devote "commensurate human and technological resources" to content moderation, and observe certain procedural guarantees such as the possibility for the user to file an appeal, especially for the most serious cases, such as the termination of an account. The bill also dictates enhanced transparency obligations and subjects digital platforms to stringent regulatory oversight. The specific requirements of this bill are expected to be aligned with the DSA (including, among other things, specific obligations for very large online platforms to assess systemic risks presented by their services and to mitigate such risks). The draft bill is set to expire when the DSA enters into force, at the latest by the end of 2023. .

Germany

Since August 2019, two legislative projects have been initiated – *inter alia* via amendments to the NetzDG, passed in 2017 – which are designed to further improve law enforcement in respect of social networks and thus to help combat illegal online content.

First, the Act to Combat Right-Wing Extremism and Hate Crime – passed by the German Bundestag on 18 June 2020 – will not only oblige social network providers to remove illegal content but, in particularly serious cases, also require them to forward the content to the Federal Criminal Police Office. This is intended to create the basis for ensuring both the fastest possible removal of content and effective criminal prosecution. Like the NetzDG, the reporting obligation also extends to criminal offences that penalise terrorist and violent extremist content (TVEC) in Germany. These include "Forming terrorist organisations" (section 129a of the Criminal Code), "Incitement of masses" (section 130 of the Criminal Code) and "Depictions of violence" (section 131 of the Criminal Code)[40].

Further amendments are planned via the Draft Act to Amend the Network Enforcement Act (NetzDGÄndG), which is currently undergoing parliamentary deliberation. Key priorities include the strengthening of user rights by creating a right to review the decisions made by social network providers on the illegality of content, and adjustments related to the assertion of rights under civil law[41].

Ireland

Ireland recently introduced the 'Online Safety and Media Regulation Bill'[42], which aims to close the legal gap in addressing harmful online content and establish a robust regulatory framework to deal with the spread of harmful online content.

Within the term 'harmful online content' is included 'content containing or comprising incitement to violence or hatred', and 'public provocation to commit a terrorist offence'.

The Bill requires transparency as a part of its online safety framework and will provide for the appointment of an Online Safety Commissioner as part of a wider Media Commission to oversee the new regulatory framework for online safety. The Commissioner will govern this new framework through binding online safety codes and robust compliance, enforcement and sanction powers. These online safety codes will deal with a wide range of issues, including:

- measures to be taken by online services to tackle the availability of harmful online content on their services;
- user complaint and/or issues handling mechanisms operated by online services;
- risk and impact assessments that may be taken by online services in relation to the availability of harmful online content on their services; and
- reporting obligations for online services.

Republic of Korea

Korea has passed several anti-terrorism laws that cover online material. Korean legislation allows the head of a related agency to request the cooperation of the head of a 'relevant institution' to eliminate, suspend and monitor suspected terrorist or violent extremist content.

In July 2016, the UN General Assembly adopted a resolution calling upon all UN Member States to develop a national plan of action to prevent violent extremism. Accordingly, the government of the Republic of Korea developed a government-wide plan for preventing violent extremism. The "National Plan of Action for Preventing Violent Extremism" was passed at the National Counter-Terrorism Committee in January 2018 and submitted to the UN. It includes plans to strengthen public-private cooperation for building a sound Internet environment and to prevent misuse of Internet and communications technologies by terrorist groups.

The Korean government is also participating in the Tech Against Terrorism Initiative led by the UN Counter-Terrorism Executive Directorate (CTED), which uses voluntary contributions for counter-terrorism and operating a Knowledge Sharing Platform for counter-terrorism. The Knowledge Sharing Platform serves as an online knowledge sharing hub that allows large enterprises to transfer their know-how about tackling the misuse of the internet by violent extremist groups to small- and medium-sized IT enterprises.

New Zealand

The New Zealand Government is continuing to progress the Films, Videos and Publications Classification (Urgent Interim Classification of Publications and Prevention of Online Harm) Amendment Bill[43]. It was introduced to Parliament on 26 May 2020 and went through first reading on 10 February 2021. Among other things, the Bill makes livestreaming of objectionable content a criminal offence. The Bill is expected to be enacted by the end of 2021.

The criminal offence of livestreaming objectionable content applies only to the individual or group livestreaming the content. It does not apply to the online content hosts that provide the online infrastructure or platform for the livestream.

Under the Bill, the Chief Censor will have powers to make immediate interim classification assessments of any publication in situations where the sudden appearance and viral distribution of objectionable content is injurious to the public good. The interim assessment will be in place until a classification decision is made or for a maximum of 20 working days, whichever is earlier. The Bill also authorises an Inspector of Publications to issue a take-down notice for objectionable online content. Such notices will be issued to an online content host and will direct the removal of a specific link to make it no longer viewable in New Zealand. Failure to comply could result in civil pecuniary penalties.

Furthermore, the Bill clarifies online content hosts' obligations in relation to objectionable material under the Films, Videos and Publications Classification Act and other types of harmful online content that falls within scope of the Harmful Digital Communications Act 2015[44] (HDCA). The HDCA aims to deter, prevent and lessen harmful digital communications, and provide victims of digital communications with a quick and efficient means of redress. Section 24 of the HDCA states that online content hosts cannot be charged under New Zealand law for hosting harmful content on their platforms if they follow certain steps when a complaint is made. The Bill makes it clear that where the online content in question is objectionable material, section 24 of the HDCA will not apply.

The Department of Internal Affairs recently established a regulatory unit to respond to reports of TVEC online, which relies on voluntary cooperation to remove TVEC. The Bill also enables future mechanisms for blocking or filtering TVEC that is deemed to be objectionable in New Zealand, should this become necessary. The Bill requires that a very clear governance and reporting system underpin any such filter.

United Kingdom

Since sketching out its plans for the regulation of platforms and content in the United Kingdom last year, the government provided further indications of what the future Online Harms Bill will look like in its Initial Consultation Response on 12 February 2020 (DCMS, 2020). DCMS Secretary of State Oliver Dowden suggested that he was considering carrying out pre-legislative scrutiny on the Online Harms Bill, thus raising suspicions that the legislation will not be introduced until the next parliamentary session (potentially as far back as 2022/23).

The Online Harms White Paper (HM Government, 2019) gives an indication of the proposed legislation's core elements:

- Services in scope of the regulation will need to ensure that illegal content is quickly removed, and that the risk of it appearing is minimised by effective systems. Companies will be required "to take particularly robust action" to tackle terrorist content and online child sexual exploitation and abuse (DCMS, 2020)
- Voluntary, interim codes of practice will provide guidance for companies on how to address online terrorist content and activities. These codes are intended to enable industry to gear up its compliance in advance of the regulator in charge of overseeing the new regulatory framework (Ofcom) becoming operational;
- A "tiered enforcement system", escalating from substantial fines, blocking of sites, criminal liability for members of a platform's senior management and ISP blocking for the most "severe" cases.
- Companies will have to demonstrate adherence to the new statutory "duty of care". The duty of care will require companies to take more responsibility for harmful content and behaviour occurring on their platforms. They will need to ensure that they have effective systems and processes in place for reducing and responding to online harm. An independent regulator will be tasked with overseeing compliance with this duty of care. The Online Harms bill suggests the following requirements for tech companies to uphold their duty of care:
    - Terms of Service (ToS) must be updated to explicitly mention which content they deem appropriate (or inappropriate) on their platforms;
    - The production of annual transparency reports

- The introduction of an easy-to-access user complaints system
- The obligation to respond to user complaints within an "appropriate timeframe" to be defined by Ofcom.

United States

The United States approach to TVEC online is guided principally by the First Amendment to the U.S. Constitution which reads, "Congress shall make no law…abridging the freedom of speech." In general, the First Amendment protects a wide range of speech—even speech that is abhorrent or offensive—and generally prohibits prior restraint or censorship of speech by the government. The government may, however, prohibit speech that is directed at inciting or producing imminent lawless action and is likely to incite or produce such action. Therefore, instead of criminalizing hateful or abhorrent speech and speech that incites violence or advocates for dangerous causes or groups, the United States has focused on prosecuting criminal activities in furtherance of violence and on promoting credible alternative narratives as the primary means to undermine and counter terrorist messaging.

A number of U.S. statutes criminalize speech-related conduct that supports violent actions, including terrorist acts. For example, under 18 U.S.C. § 373, it is a crime to solicit, command, induce, or otherwise endeavour to persuade another person to engage in a felony involving the threatened, attempted, or actual use of physical force against another person or property, in violation of the laws of the United States.

Additionally, the material support to foreign terrorist organisations statute, 18 U.S.C. § 2339B, applies to actions knowingly made in support or under the direction of, or in coordination with, designated foreign terrorist organisations that the actor knows to be terrorist organisations.

Under U.S. law, online service providers are generally protected from liability for the speech of their users, and are protected from liability for their content moderation decisions, except in limited circumstances, including for violations of federal criminal law (see Section 230 of the Communications Act of 1934). The U.S. intermediary liability framework facilitates the ability of online service providers to moderate the use of their platforms for types of speech that could not be banned by the government.

Annex A – List of the Top 50 Services

Rank	Name of Service (parent company)	Monthly active users, user accounts or unique visitors (millions)	Type of Service	Issues TVEC transparency reports	Provided feedback / comments on its profile
1	Facebook (Facebook, Inc.)	2,603 (as of July 2020) (Kemp, More than Half of the People on Earth now Use Social Media, 2020)	Social networking and video streaming platform	Y	Y
2	YouTube (Alphabet, Inc.)	2,000 (as of July 2020) (Kemp, More than Half of the People on Earth now Use Social Media, 2020)	Video streaming platform	Y	Y
3	WhatsApp (Facebook, Inc.)	2,000 (as of July 2020) (Kemp, More than Half of the People on Earth now Use Social Media, 2020)	Messaging app	N	Y
4	Facebook Messenger (Facebook, Inc.)	1,300 (as of July 2020) (Kemp, More than Half of the People on Earth now Use Social Media, 2020)	Messaging app	N	Y
5	iMessage/FaceTime (Apple, Inc)	1,300 (as of January 2019) (Elmer-	Messaging and video chat apps	N	N

		Dewitt, 2019)			
6	Weixin/WeChat (Tencent Holdings Ltd.)	1,203 (as of July 2020) (Kemp, More than Half of the People on Earth now Use Social Media, 2020)	Social networking/content sharing/messaging platform	N	N
7	Instagram (Facebook, Inc.)	1,082 (as of July 2020) (Kemp, More than Half of the People on Earth now Use Social Media, 2020)	Social networking platform	Y	Y
8	Tik Tok (ByteDance Technology Co.)	800 (as of July 2020) (Kemp, More than Half of the People on Earth now Use Social Media, 2020)	Short video app	Y	N
9	QQ (Tencent Holdings Ltd.)	694 (as of July 2020) (Kemp, More than Half of the People on Earth now Use Social Media, 2020)	Instant messaging and web portal site	N	N
10	Youku Tudou (Alibaba Group Holding Limited)	580 (as of August 2019) (Youku Tudou Inc. (NYSE: YOKU), n.d.)	Video streaming platform (user-generated and syndicated content)	N	N
11	Weibo (Sina Corp.)	550 (as of July 2020) (Kemp, More than Half of the People on Earth now Use Social Media, 2020)	Social networking platform	N	N

12	QZone (Tencent Holdings Ltd.)	517 (as of July 2020) (Kemp, More than Half of the People on Earth now Use Social Media, 2020)	Social networking platform	N	N
13	iQIYI (Baidu, Inc.)	476 (as of December 2019) (Statista, 2019)	Video streaming platform (user-generated and syndicated content)	N	N
14	Reddit (Reddit, Inc.)	430 (as of July 2020) (Kemp, More than Half of the People on Earth now Use Social Media, 2020)	Social news aggregation, web content ranking and discussion website	Y	Y
15	Kuaishou (Beijing Kuaishou Technology Co., Ltd)	400 (as of July 2020) (Kemp, More than Half of the People on Earth now Use Social Media, 2020)	Short video app	N	N
16	Telegram (Telegram Messenger LLP)	400 (as of April 2020) (Singh, 2020)	Messaging app	N	N
17	Snapchat (Snap, Inc.)	397 (as of July 2020) (Kemp, More than Half of the People on Earth now Use Social Media, 2020)	Social networking platform	N	Y
18	Pinterest (Pinterest, Inc.)	367 (as of July 2020) (Kemp, More than Half of the People on Earth now Use Social Media, 2020)	Social networking platform	N	N

19	Twitter (Twitter, Inc.)	326 (as of July 2020) (Kemp, More than Half of the People on Earth now Use Social Media, 2020)	Short messages-focused social networking platform	Y	N
20	Douban (Information Technology Company, Inc.)	320 (as of July 2019) (Kemp, Digital 2019: Q3 Global Digital Statshot, 2019)	Social networking platform	N	N
21	LinkedIn (Microsoft, Inc.)	310 (as of July 2019) (Kemp, Digital 2019: Q3 Global Digital Statshot, 2019)	Jobs-focused social networking platform	N	Y
22	Baidu Tieba (Baidu, Inc.)	300 (as of March 2020) (Marketing to China, 2020)	Online communications platform	N	N
23	Skype (Microsoft, Inc.)	300 (as of June 2019) (Perez, Skype publicly launches screen sharing on iOS and Android, 2019)	Video chat and voice calls app	Y	Y
24	Quora (Quora, Inc.)	300 (as of September 2018) (Marketing Land, 2018)	Question-and-answer website	N	N
25	Xigua (ByteDance Technology Co.)	270 (as of December 2019) (Chen, 2020)	Short video streaming app	N	N
26	Viber (Rakuten, Inc.)	260 (as of July 2019) (Kemp, Digital 2019: Q3	Messaging app	N	Y

		Global Digital Statshot, 2019)			
27	Discord (Discord, Inc.)	250 (as of July 2019 (Kemp, Digital 2019: Q3 Global Digital Statshot, 2019)	Chat platform	Y	N
28	Vimeo (Vimeo, Inc.)	240 (as of September 2018) (Bicknell, 2018)	Video streaming app	N	N
29	IMO (PageBites, Inc.)	211 (as of April 2019) (YY Inc. - IR Site, 2019)	Video chat and voice calls app	N	N
30	LINE (Line Corporation)	194 (as of January 2019) (Kemp, Digital 2019: Global Digitial Overview, 2019)	Messaging app	N	N
31	Huoshan (ByteDance Technology Co.)	170 (as of December 2019) (Chen, 2020)	Short video streaming app	N	N
32	Ask.fm (IAC [InterActiveCorp])	160 (as of April 2020) (Kallas, 2020)	Social networking platform	N	Y
33	YY Live/Huya (YY, Inc.)	157 (as of November 2019) (Yahoo! Finance, 2019)	Livestreaming platform	N	N
34	Twitch (Amazon.com, Inc.)	140 (as of July 2020) (Iqbal, 2020)	Livestreaming platform	Y	Y
35	Tumblr (Automattic, Inc.).	115 (as of April 2020) (Kallas, 2020)	Microblogging and social networking platform	N	N

36	Flickr (SmugMug, Inc.)	112 (as of April 2020) (Kallas, 2020)	Image and video hosting service	N	N
37	VK (Mail.Ru Group)	97 (as of April 2020) (Kallas, 2020)	Social networking platform	N	Y
38	Medium (A Medium Corporation.)	86 (as of August 2018) (Wickey, 2018)	Online publishing platform	N	Y
39	Odnoklassniki (Mail.Ru Group)	71 (as of April 2020) (Kallas, 2020)	Social networking platform	N	N
40	Haokan (Baidu, Inc.)	69 (as of June 2019) (Chen, 2020)	Short video streaming app	N	N
41	Smule (Smule, Inc.)	52 (as of July 2018) (Solsman, 2018)	User-generated music-video sharing platform	N	N
42	KaKao Talk (Daum Kakao Corporation)	50 (as of June 2019) (Statista, 2019)	Messaging app	N	Y
43	Deviantart (DeviantArt, Inc.)	45 (as of 2016) (DeviantArt Media Kit, n.d.)	Online artwork, videography and photography platform	N	N
44	Meetup (WeWork Companies, Inc.)	35 (as of April 2020) (Kallas, 2020)	Interest-based social networking platform	N	N
45	4chan (4chan Community Support LLC)	22 (as of August 2019) (4chan, n.d.)	Content sharing platform	N	N

Monthly active user (MAU) data are unavailable for certain other online content-sharing services that terrorists and violent extremists have used, yet the metrics that are available suggest that they should be included in the top 50 list. The table therefore continues below with five more services, but without ranks because metrics other than MAU indicate their significance, so a proper comparison with the services above was not possible. In any event, for purposes of this report, the overall composition of the group of 50 is more important than the individual rankings.

Name of Service (parent company)	Indicative Global Market Share	Type of market/Service	Transparency report on terrorist/violent extremist content	Provided feedback / comments on its profile

Google Drive (Alphabet, Inc.)	34.35% (as of October 2019) (Datanyze, 2020)	Cloud-based file sharing	N	Y
Dropbox (Dropbox, Inc.)	21.23% (as of October 2019) (Datanyze, 2020)	Cloud-based file sharing	N	Y
Microsoft OneDrive (Microsoft, Inc.)	12.07% (as of October 2019) (Datanyze, 2020)	Cloud-based file sharing	Y	Y

Name of Service (parent company)	**Indicative Global Market Share or monthly pageviews**	**Type of market/Service**	**Transparency report on terrorist/violent extremist content**	**Provided feedback / comments on its profile**
Wordpress.com (Automattic, Inc.)	60% (as of April 2019) (Kinsta, 2011-2019)	Content management system	Y	Y
Wikipedia (Wikimedia Foundation)	18 billion pageviews per month (as of January 2016) (Pew Research Center, 2016); 10th most visited website worldwide (Alexa, 2019)	Online encyclopaedia	N	N

Annex B - Profiles of the Top 50 Services

1. Facebook[45]

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There is no specific definition of TVEC. However, Facebook is one of the few Services with a well-developed definition of terrorism and related terms. In the section of Facebook's Community Standards entitled 'Dangerous Individuals and Organisations (Facebook), Facebook states that any organisations or individuals that proclaim a violent mission or are engaged in violence cannot have a presence on Facebook. Such organisations or individuals are defined to include those involved in: ● Terrorist activity ● Organised hate ● Mass murder (including attempts) or multiple murder ● Human trafficking ● Organized violence or criminal activity Content that expresses support or praise for groups, leaders or individuals involved in these activities is removed. Also, the following people (whether living or deceased) and groups cannot maintain a presence (for example, have an account, Page or group) on Facebook: terrorist organisations, terrorists, hate organisations (and their leaders and prominent members) and mass and multiple murderers. Terrorist organisations and terrorists include any non-state actor that: ● Engages in, advocates or lends substantial support to purposive and planned acts of violence, ● Which causes or attempts to cause death, injury or serious harm to civilians, or any other person not taking direct part in the hostilities in a situation of armed conflict, and/or significant damage to property linked to death, serious injury or serious harm to civilians ● With the intent to coerce, intimidate and/or influence a civilian population, government or international organisation ● In order to achieve a political, religious or ideological aim.

	A hate organisation is defined as any association of three or more people that is organised under a name, sign or symbol and that has an ideology, statements or physical actions that attack individuals based on characteristics, including race, religious affiliation, nationality, ethnicity, gender, sex, sexual orientation, serious disease or disability. A homicide is considered to be a mass murder if it results in three or more deaths in one incident. Any individual who has committed two or more murders over multiple incidents or locations is deemed a multiple murderer. Facebook prohibits any symbols that represent any of the above organisations or individuals, unless they are shared with context that condemns or neutrally discusses the content. Content that praises any of the above organisations or individuals or any acts committed by them is prohibited. Also, Facebook does not allow coordination of support for any of the above organisations or individuals or any acts committed by them. Further, Facebook prohibits content that represents or supports in any way events that it designates as terrorist attacks, hate crimes, or mass shootings. Lastly, in the section titled 'Violence and Incitement' of Facebook's Community Standards (Facebook), Facebook states that it removes language that incites or facilitates serious violence. In particular, users cannot post: ● Threats that could lead to death (and other forms of high-severity violence) of any target(s), where threat is defined as any of the following: ○ Statements of intent to commit high-severity violence ○ Calls for high-severity violence including content where no target is specified but a symbol represents the target and/or includes a visual of an armament to represent violence; or ○ Statements advocating for high-severity violence; or ○ Aspirational or conditional statements to commit high-severity violence ● Content that asks or offers services for hire to kill others (for example, hitmen, mercenaries, assassins) or advocates for the use of a hitman, mercenary or assassin against a target.

	● Admissions, statements of intent or advocacy, calls to action or aspirational or conditional statements to kidnap a target. ● Threats that lead to serious injury (mid-severity violence) towards private individuals, minor public figures, vulnerable persons or vulnerable groups, where threat is defined as any of the following: o Statements of intent to commit violence o Statements advocating violence; or o Calls for mid-severity violence including content where no target is specified but a symbol represents the target; or o Aspirational or conditional statements to commit violence; or o Content about other target(s) apart from private individuals, minor public figures, vulnerable persons or vulnerable groups and any credible: ▪ Statements of intent to commit violence; ▪ Calls for action of violence; ▪ Statements advocating for violence; or ▪ Aspirational or conditional statements to commit violence ● Threats that lead to physical harm (or other forms of lower-severity violence) towards private individuals (self-reporting required) or minor public figures, where threat is defined as any of the following: o Statements of intent o calls for action o advocating, aspirational, or conditional statements to commit low-severity violence ● Imagery of private individuals or minor public figures that has been manipulated to include threats of violence either in text or pictorially (adding bullseye, dart, gun to head etc.) ● Any content created for the express purpose of outing an individual as a member of a designated and recognisable at-risk group ● Instructions on how to make or use weapons if there is evidence of a goal to seriously injure or kill people, through: o Language explicitly stating that goal, or

	○ photos or videos that show or simulate the end result (serious injury or death) as part of the instruction, ○ unless the aforementioned content is shared as part of recreational self-defence, for military training purposes, commercial video games or news coverage (posted by Page or with news logo) ● Providing instructions on how to make or use explosives, unless there is clear context that the content is for a non-violent purpose (for example, part of commercial video games, clear scientific/educational purpose, fireworks or specifically for fishing) ● Any content containing statements of intent, calls for action or advocating for high or mid-severity violence due to voting, voter registration or the outcome of an election ● Misinformation that contributes to imminent violence or physical harm; and ● Calls to action, statements of intent to bring armaments to locations, including but not limited to places of worship, or encouraging others to do the same.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.facebook.com/communitystandards/.
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	Yes, available at https://about.fb.com/news/2019/05/protecting-live-from-abuse/. In particular, Facebook applies a 'one strike' policy to prohibited livestreamed content, meaning that anyone who violates Facebook's 'most serious policies' will be restricted from using Live for set periods of time, for example 30 days, starting on their first offense.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	Facebook removes content from the platform when content violates its Community Standards.
4.1 Notifications of removals or other enforcement decisions	After the content removal, the person who posted the content is notified and given the option to request a review or accept the decision (Facebook).

4.2 Appeal processes against removals or other enforcement decisions	If the user requests a review, the content is resubmitted for another review. The content is not visible to other people on Facebook while under review. Reviewers do not know that the post has been reviewed previously. Based on the wording of the document 'Understanding the Community Standards Report' (Facebook), it seems that the review is done by a single person. If the reviewer agrees with the original decision, the content remains off Facebook. However, if the reviewer disagrees with the initial review and decides it should not have been removed, the content will go to a third reviewer. This reviewer's decision will determine whether the content is allowed on Facebook or not.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Facebook detects violations to its policies, including its policy on Terrorist Propaganda, through a combination of technology, reports from users and reviews by its teams (Facebook). In particular, Facebook uses artificial intelligence (AI) to combat terrorism, including techniques such as image matching, language understanding, removal of terrorist clusters and cross-platform collaboration with other Facebook-owned platforms (i.e. with WhatsApp and Instagram). Around three years ago, Facebook started using machine learning to assess Facebook posts that may signal support for ISIS or al-Qaeda (Facebook, 2018). Since then Facebook expanded these techniques to detect and remove content related to other terrorist groups and organized hate. Facebook is now able to detect text embedded in images and videos in order to understand its full context, and it has built media matching technology to find content that is identical or near-identical to photos, videos, text and even audio that Facebook has already removed. When Facebook started detecting hate organisations, it focused on groups that posed the greatest threat of violence at that time. Now it has expanded to detect more groups tied to different hate-based and violent extremist ideologies and using different languages. In addition to building new tools, Facebook has also adapted strategies from its counterterrorism work, such as leveraging off-platform signals to identify dangerous content on Facebook, and implementing procedures to audit the accuracy of its AI's decisions over time (Facebook, 2020). Facebook has reported that the video of the attack in Christchurch did not prompt its automatic detection systems because it did not have enough content depicting first-person footage of violent events to effectively train its machine learning technology. Accordingly, Facebook started working with government and law enforcement officials in the US

	and UK to obtain camera footage from their firearms training programs – providing a valuable source of data to train its systems. With this initiative, Facebook aims to improve its detection of real-world, first-person footage of violent events and avoid incorrectly detecting other types of footage such as fictional content from movies or video games. (Facebook, 2019) Facebook notes that AI cannot catch everything. Therefore, user reports play a fundamental role in the detection of objectionable content, allowing Facebook to identify new concerns quickly, as well as to improve the signals used in its technology to detect policy violations (Facebook). Lastly, Facebook has a dedicated 'Community Operations Team' that reviews content and additional context to determine whether it violates its policies. This team includes experts in the field of terrorism. This team reviews reports 24 hours a day, 7 days a week, and the vast majority of reports are reviewed within 24 hours (Facebook). According to Facebook, whether identified by its technology or reported by users, a flagged potential violation becomes a report in its system. Facebook prioritizes safety-related reports, including material related to terrorism and suicide. Facebook then uses technology, human review or a combination of the two to determine whether a piece of content violates its policies. If the content is routed to its human review team, then they use Facebook's policies and a step-by-step process to help them make decisions accurately and consistently for the appropriate violation type. Facebook also provides its reviewers with tools to review the reported content and the available context required to identify the concern and determine whether a piece of content violates a standard (Facebook). The marginal economic costs of using AI tools to identify TVEC are probably very low (although fixed costs may be substantial), whereas the marginal economic costs of using human moderators to this end are probably relatively high. Facebook is a founding member of GIFCT and participates in its Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	The consequences for breaching Facebook's Community Standards vary depending on the severity of the breach and a person's history on the platform. Prohibited content may be removed. In addition, Facebook may warn someone after a first breach, but if the user continues to breach Facebook's policies, Facebook may restrict the user's ability to post on

	Facebook or disable their profile. Facebook may also notify law enforcement when it believes that there is a genuine risk of physical harm or a direct threat to public safety.
7. Does the service issue transparency reports (TRs) specifically on content related to terrorism and/or violent extremism?	Yes (Facebook, 2017-2020). Facebook issues transparency reports on the enforcement of its Community Standards, in which one section is about 'Dangerous Organisations: Terrorism and Organised Hate', while another is about 'Violence and Graphic Content'. Note that Facebook states that it does not tolerate any content that praises, endorses or represents individuals or groups engaging in terrorist activity or organised hate. Facebook enforces this standard as applied to terrorist activities and groups both regionally and globally. Since November 2019, its terrorist propaganda TRs measure the actions Facebook takes against all terrorist organisations, rather than focusing just on propaganda related to ISIS, al-Qaeda and their affiliate groups (Facebook, 2020).
8. What information/fields of data are included in the TRs?	In the latest report issued in November 2020, the following five fields of information were included in both the 'Dangerous Organisations: Terrorism and Organised Hate' section and the 'Violence and Graphic Content' section: - *Prevalence (How prevalent were terrorism and violence and graphic content violations on Facebook?)* The prevalence metric is the percentage of views that included terrorism and violence and graphic content violations. For example, Facebook estimated an upper limit of 0.05% of views of content that violated its terrorism standards in Q2 2020. That means that out of every 10,000 views for the terrorism policy on Facebook, no more than 5 of those views contained content that violated that policy. (The figures refer to final determinations, not content that was initially flagged as a possible violation but may have been subsequently determined to be permissible.) - *Content actioned (How much content did Facebook take action on?)* Facebook indicates that a piece of content can be 'any number of things', (Facebook) including a post, photo, video or comment. Taking action may include removing a piece of content from Facebook, covering photos or videos that may be disturbing to some audiences with a warning, or disabling accounts. Content actioned is the total number of pieces of content that Facebook took action on during a given reporting period because it violated its

	community standards (in this case the terrorism and violence and graphic content policies). - *Proactive rate (Of the violating content actioned, how much did Facebook find before users reported it?)* This metric shows the percentage of content actioned for dangerous organisations and violence and graphic content that Facebook found and flagged before users reported it. It counts detections made by both Facebook's AI tools and human reviewers. - *Appealed Content (How much of the content Facebook actioned did people appeal?)* This metric counts the number of pieces of content actioned for which people requested another review during the reporting period. - *Restored Content (How much content did Facebook restore after removing it?)* Restored content is the number of pieces of content that Facebook restored during the reporting period after previously actioning it. Facebook also includes recent trends regarding content actioned for organised hate and terrorism. For example, its last transparency report notes that content actioned for organised hate decreased from 4.7 million pieces of content in Q1 2020 to 4 million in Q2 2020, and content actioned for terrorism increased from 6.3 million pieces of content in Q1 2020 to 8.7 million in Q2 2020.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	- *Prevalence.* The prevalence metric is the estimated number of views of violating content, divided by the estimated number of total content views on Facebook, per reporting period. For example, if the prevalence of dangerous organisations is 0.18% to 0.20%, that means of every 10,000 content views, 18 to 20 on average were of content that violated Facebook's standards for dangerous organisations. The prevalence metric provides an indication of how often prohibited content is seen, rather than the total amount of such content published. Prevalence is estimated based on samples of content across different areas of Facebook, such as Groups and News Feeds. For terrorism violations, in particular, Facebook only estimates the upper limit, which means that Facebook is 'confident that the prevalence of violating views is below that limit.' (Facebook). Facebook elaborates on the prevalence methodology in 'Measuring

	Prevalence of Violating Content on Facebook' (Facebook, 2019). - *Content actioned.* Content actioned is the total number of pieces of content that Facebook took action on during a given reporting period because it violated its content policies. Facebook does not count those scenarios where it escalates content to law enforcement. This metric includes both content Facebook actioned after someone reported it and content that Facebook found proactively. Content on Facebook and Messenger are included in this metric. - *Proactive rate.* This metric is calculated as: the number of pieces of content actioned that Facebook found and flagged before users reported them, divided by the total number of pieces of content actioned. Content on Facebook and Messenger are included in this metric. - *Appealed Content.* This metric counts the number of pieces of content actioned for which people requested another review during the reporting period. Content on Facebook and Messenger are included in this metric. Facebook observes that this metric shows the number of pieces of content that were appealed within the quarter, whereas restored content (see below) counts the content restored within the quarter. Because some appealed content may be restored in the following quarter, and some restored content was appealed in a previous quarter, these metrics cannot be directly compared. - *Restored content.* To arrive at this metric, Facebook counts the number of pieces of content that it restored during the reporting period after previously actioning it. Facebook may restore content either when a decision to remove is appealed or when Facebook discovers a reason to restore the content. Only Facebook content is included in this metric.
10. Frequency/timing with which TRs are issued	As from August 2020, Facebook publishes its TRs on a quarterly basis. Its last report covers Q2 2020. Currently, there is available data from Q4 2017 to Q2 2020.
11. Has this service been used to post TVEC?	Yes. See above sections 7-9.

2. YouTube

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There is no specific definition of TVEC. However, YouTube's Community Guidelines contain a number of clarifications that are relevant to terrorist and violent extremist content. The policy on Violent Criminal Organisations, for example, states that content intended to praise, promote, or aid violent criminal organisations is not allowed on YouTube. In addition, such organizations are banned from YouTube for any purpose, including recruitment. The Guidelines neither contain nor refer to a list of such organisations, though. Nevertheless, the policy prohibits the following types of content: • Content produced by violent criminal or terrorist organisations • Content praising or memorialising prominent terrorist or criminal figures in order to encourage others to carry out acts of violence • Content praising or justifying violent acts carried out by violent criminal or terrorist organisations • Content aimed at recruiting new members to violent criminal or terrorist organisations • Content depicting hostages or posted with the intent to solicit, threaten, or intimidate on behalf of a violent criminal or terrorist organisation • Content that depicts the insignia, logos, or symbols of violent criminal or terrorist organisations in order to praise or promote them. If content related to terrorism or crime is posted for an educational, documentary, scientific, or artistic purpose, enough information in the video or audio must be included so viewers understand the context. The policy on Violent Criminal Organisations also gives the following examples of content that is not allowed on YouTube: • Raw and unmodified reuploads of content created by terrorist or criminal organisations • Celebrating terrorist leaders or their crimes in songs or memorials • Celebrating terrorist or criminal organisations in songs or memorials • Content directing users to sites that espouse terrorist ideology, are used to disseminate prohibited content, or are used for recruitment. • Video game content which has been developed or modified ('modded') to glorify a violent event, its perpetrators, or support violent criminal or terrorist organisations.

	Moreover, YouTube's violent or graphic content policies prohibits violent or gory content intended to shock or disgust viewers, or content encouraging others to commit violent acts. In particular, YouTube prohibits the following types of content: • Inciting others to commit violent acts against individuals or a defined group of people • Footage, audio or imagery involving road accidents, natural disasters, war aftermath, terrorist attack aftermath, street fights, physical attacks, sexual assaults, immolation, torture, corpses, protests or riots, robberies, medical procedures or other such scenarios with the intent to shock or disgust viewers. In turn, YouTube's policy on hate speech bans content promoting violence or hatred against individuals or groups based on any of the following attributes: Age, Caste, Disability, Ethnicity, Gender Identity, Nationality, Race, Immigration Status, Religion, Sex/Gender, Sexual Orientation, Victims of a major violent event and their kin, and Veteran Status. Content that encourages violence against individuals or groups based on any of on the attributes noted above, or that incites hatred against individuals or groups based on any of the attributes noted above, is prohibited. Among the examples provided of content that falls within this category is praising or glorifying violence against individuals or groups based on the attributes noted above. In June 2019 YouTube updated its hate speech policy to specifically prohibit videos alleging that a group is superior in order to justify discrimination, segregation or exclusion based on attributes like age, gender, race, caste, religion, sexual orientation or veteran status. YouTube also announced that it will remove content denying that well-documented violent events took place (Google, Youtube, 2019). Lastly, the policy on harmful or dangerous content bans instructions to kill or harm. This means showing viewers how to perform activities meant to kill or maim others, such as providing instructions on how to build a bomb meant to injure or kill people. Also prohibited is content about violent events if it promotes or glorifies violent tragedies such as school shootings.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	YouTube's Community Guidelines are available at https://www.youtube.com/about/policies/#community-guidelines Guidelines on Violent Criminal Organisations are available at https://support.google.com/youtube/answer/9229472?hl=en&ref_topic=9282436 Guidelines on violent or graphic content are available at https://support.google.com/youtube/answer/2802008?hl=en-GB&ref_topic=9282436 Guidelines on hate speech are available at https://support.google.com/youtube/answer/2801939?hl=en

	Guidelines on harmful or dangerous content are available at https://support.google.com/youtube/answer/2801964?hl=en&ref_topic=9282436
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No. YouTube's Community Guidelines apply to videos, video descriptions, comments, live streams and any other YouTube product or feature.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	If content violates any of YouTube's content policies, YouTube removes the content.
4.1 Notifications of removals or other enforcement decisions	The content removal is notified to users via email, desktop or mobile notifications, and an alert in their channel settings (Google/Youtube, 2020). If the content removal results in a 'strike' (see below section 6), YouTube informs the user: • What content was removed • Which policies it violated • How the strike affects the user's channel • What the user can do next
4.2 Appeal processes against removals or other enforcement decisions	When users receive a strike, and they believe YouTube made a mistake, they can appeal the strike (Google, Youtube, 2020). YouTube informs users about the result of the appeal via email. The result may be any of the following: • If YouTube finds that the content followed YouTube's Community Guidelines, YouTube reinstates it and removes the strike from the user's channel. If the user appeals a warning (see below section 6) and the appeal is granted, the next offense will result in a warning. • If YouTube finds that the content followed YouTube's Community Guidelines, but is not appropriate for all audiences, an age-restriction is applied. If the content is a video, it will not be visible to users who are signed out, are under 18 years of age, or have Restricted Mode (Google, Youtube, 2020) turned on. If the content is a custom thumbnail, it will be removed. • If YouTube finds that the content was in violation of YouTube's Community Guidelines, the strike will stay

	and the video will remain off the platform. There is no additional penalty for appeals that are rejected. Users may appeal each strike only once.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	YouTube provides its users with tools to report content that violates its Community Guidelines (Google, Youtube, 2020). YouTube has also developed automated systems that aid in the detection of content that may violate its policies. When its automated systems flag potentially problematic content, human reviewers then verify whether it indeed violates company policies. If it does, the content is removed and is used to train YouTube's automated systems to perform better in the future. With respect to the automated systems that detect extremist content (an undefined term) in particular, YouTube's staff have manually reviewed over two million videos to provide training examples. In addition, YouTube invests in a network of over 180 academics, government partners and NGOs who bring expertise to the platform's enforcement systems, including through YouTube's Trusted Flagger programme. (Google, Youtube, 2020)[46] In the context of violent extremism, this includes the International Centre for the Study of Radicalisation at King's College, London (The International Centre for the Study of Radicalisation (ICSR), 2020), the Institute for Strategic Dialogue (ISDGlobal, n.d.), the Wahid Institute in Indonesia and government agencies focused on counterterrorism. Participants in the Trusted Flagger programme receive training in enforcing YouTube's Community Guidelines, and because their flags have a higher action rate than the average user, YouTube prioritises them for review. Otherwise, content flagged by Trusted Flaggers is subject to the same policies as content flagged by any other user and is reviewed by teams that are trained to make decisions on whether content violates YouTube's Community Guidelines. Individual users, government agencies, and NGOs are eligible for participation in the YouTube Trusted Flagger programme. Participants must be committed to frequently flagging content that may violate YouTube's Community Guidelines and be open to ongoing discussion and feedback on various YouTube content areas. YouTube notes that hate speech is a complex policy area to enforce at scale, as decisions require nuanced understanding of local languages and contexts. For consistent enforcement of its hate speech policy, YouTube has expanded its review team's linguistic and subject matter expertise. YouTube also deploys machine learning to better detect potentially hateful content to send for human review, applying lessons from its enforcement against other types of content, like violent extremism (Google, Youtube, n.d.). The marginal economic costs of using automated tools to identify TVEC are probably very low (although fixed costs may be substantial), whereas the marginal economic costs of using human moderators to this end are probably relatively high.

	YouTube is a founding member of GIFCT and participates in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	The first time a user posts content that violates YouTube's Community Guidelines, he or she receives a warning with no penalty to their channel. For subsequent violations, YouTube issues a 'strike' against the user's channel. The channel is terminated if the user receives 3 strikes within a 90-day period. When the first strike is issued, the user cannot do any of the following for one week: • Upload videos, live streams, or stories • Create custom thumbnails or Community posts • Created, edit, or add collaborators to playlists • Add or remove playlists from the watch page using the "Save" button Full privileges are restored automatically after the 1-week period, but the strike will remain on the user's channel for 90 days. If the user gets a second strike within 90-days of the first strike, the user will not be able to post content for two weeks. If there are no further issues, full privileges are restored automatically after the 2-week period, but each strike expires 90 days from the time it was issued. Three strikes in the same 90-day period will result in the user's channel being permanently removed from YouTube (Google, YouTube, n.d.). Beyond the three strikes system, a YouTube channel will be terminated if it has a single case of severe abuse (such as predatory behaviour) or is determined to be wholly dedicated to violating YouTube's guidelines (as is often the case with spam accounts). When a channel is terminated, all of its videos are removed. Content that does not violate YouTube's policies but is close to meeting the criteria for removal and could be offensive to some viewers may have some features disabled. The content will remain available on YouTube, but the watch page will no longer have comments, suggested videos or likes, and will be placed behind a warning message. These videos are also not eligible for ads. Having features disabled will not add a strike to the video owner's channel (Google, YouTube, n.d.). YouTube notifies decisions to disable features via email. Users can appeal this decision.

7. Does the service issue transparency reports (TRs) on TVEC?	Yes (Google, n.d.). YouTube issues transparency reports on the enforcement of its Community Guidelines. One section of these reports is about 'Violent Extremism' (Google, YouTube, n.d.). The last TR specifies that content that violates YouTube's policies against violent extremism includes material produced by government-listed foreign terrorist organisations (YouTube does not specify which government(s) it is referring to, though). The TR also specifies that YouTube strictly prohibits content that promotes terrorism, such as content that glorifies terrorist acts or incites violence. In addition, the TR states that content produced by violent extremist groups that are not government-listed foreign terrorist organisations is often covered by YouTube's policies against posting hateful or violent or graphic content (see Section 1 above), including content that is primarily intended to be shocking, sensational or gratuitous.
8. What information/fields of data are included in the TRs?	YouTube discloses • the number of content removal requests by governments based on six categories (national security, defamation, regulated goods and services, privacy and security, copyrights and 'all others') (Google, 2010-2020); • the number of channels removed, separated by ground of removal (amongst which are the promotion of violence and violent extremism); • the number of videos removed by source of first detection (automated flagging, individual trusted flagger, users, NGOs and government agencies); • the percentage of videos first flagged through automated flagging systems, with and without views, i.e. the percentage of removals that occurred before the videos received any views versus those that occurred after the videos received some views; • the number and percentage of human flags, by flagging reason (including the promotion of terrorism). YouTube notes that a video may be flagged multiple times for multiple reasons, and that flagging it does not necessarily result in removal. Human-flagged videos are removed for violations of Community Guidelines once a trained reviewer confirms a policy violation (Google, Youtube, 2017-2020). • the total number of appeals that YouTube received for videos removed due to a community violation per quarter, and the total number of videos that YouTube reinstated due to an appeal after being removed for a community guidelines violation per quarter. • the percentage and number of videos removed, by removal reason (including under YouTube's violent extremism policy and hate speech policy) (Google, YouTube, n.d.); • the number of comments removed, by removal reason (including under YouTube's violent extremism policy and hate speech policy); and • the percentage of removed comments by source of first detection (automated flagging and human flagging).

	YouTube's transparency report features a section titled 'featured policies', which include the total number of videos removed for violation of its Violent Extremism and Hate Speech policies.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	No information is provided.
10. Frequency/timing with which TRs are issued	On a quarterly basis. Last TR covers Q2 2020.
11. Has this service been used to post TVEC?	Yes. See above sections 7-8.

3. WhatsApp

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	WhatsApp's ToS do not define TVEC. However, in the section titled 'Safety and Security' in WhatsApp's ToS states that WhatsApp works to protect the safety and security of WhatsApp by appropriately 'dealing with abusive people and activity' and violations of its Terms. It is possible that the concept 'abusive people and activity' encompasses users disseminating TVEC, although this is not stated explicitly. 'Abusive people and activity' is not defined. The ToS also state that WhatsApp prohibits misuse of its services, 'harmful conduct towards others', and violations of its Terms and policies. WhatsApp notes that users must access and use its services only for 'legal, authorised, and acceptable purposes', which includes not using its services in ways that "are illegal, obscene, defamatory, threatening, intimidating, harassing, hateful, racially or ethnically offensive, or instigate or encourage conduct that would be illegal or otherwise inappropriate, including promoting violent crimes."
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.whatsapp.com/legal/#terms-of-service
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No. WhatsApp does not have joinable live streamed content.

4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	WhatsApp broadly states that it may modify, suspend, or terminate a user's access to or use of its services at any time for suspicious or unlawful conduct, or if it reasonably believes that the user is violating its Terms or creating harm or risk for users or other people. No appeal processes are specified. However, if a user believes that his or her account was terminated or suspended by mistake, the user can contact WhatsApp at support@whatsapp.com.
4.1 Notifications of removals or other enforcement decisions	If a number is banned, the user receives a notification, as explained at https://faq.whatsapp.com/general/account-and-profile/seeing-the-message-your-phone-number-is-banned-from-using-whatsapp-contact-support-for-help/?lang=en
4.2 Appeal processes against removals or other enforcement decisions	No appeal processes are specified. However, if a user believes that his or her account was terminated or suspended by mistake, the user can contact WhatsApp at support@whatsapp.com.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	WhatsApp states that it maintains advanced machine learning technology to evaluate group information including names, profile photos, and group descriptions to improve its ability to detect and remove 'abusive people and activity' that may harm WhatsApp's community and the safety and security of its services. Also, users can report any content they may deem problematic, and WhatsApp's moderators review those reports to take appropriate action. WhatsApp also states that it prevents chat groups from maintaining certain representations, such as using particular group names, in order to meet its obligations prescribed by U.S. law related to designated terrorist organizations. The marginal economic costs of using automated tools to identify TVEC are probably very low (although fixed costs may be substantial), whereas the marginal economic costs of using human moderators to this end are probably relatively high. WhatsApp is a member of the GIFCT.
6. Sanctions/consequences in case of breaches of ToS or Community Guidelines/Standards	If a user violates WhatsApp's ToS or policies, WhatsApp may take action with respect to the user's account, including disabling or suspending it. If WhatsApp does so, the user must not create another account without WhatsApp's permission. If WhatsApp has taken action to end a group, participants will no longer be able to send messages

	to that group. In addition, WhatsApp states that it may ban administrators of such groups from using WhatsApp altogether. WhatsApp also notes that if it becomes aware of 'abusive people or activity', it will take appropriate action by removing such people or activity or contacting law enforcement.
7. Does the service issue transparency reports (TRs) on TVEC	Not yet, but 'public data transparency' is a condition of membership in GIFCT, so WhatsApp may be expected to do so in the near future.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Yes. For example, after the Christchurch shootings, two far-right violent extremists reportedly were part of a WhatsApp group called 'Christian White Militia' and published statements encouraging terrorism in March 2019 (Dearden, 2019).

4. Facebook Messenger

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There is no specific definition of TVEC. However, Facebook is one of the few Services with a well-developed definition of terrorism and related terms in Facebook's Community Standards. Under section 3 of its Terms of Service and 1.3 of its Developer Policy, Facebook's Community Standards also apply to Messenger for content generated by users or messaging bots. In the section of Facebook's Community Standards, entitled 'Dangerous Individuals and Organisations (Facebook, n.d.[1]), Facebook states that any organisations or individuals that proclaim a violent mission or are engaged in violence cannot have a presence on Facebook products. Such organisations or individuals are defined to include those involved in: • Terrorist activity • Organised hate

	• Mass murder (including attempts) or multiple murder • Human trafficking • Organized violence or criminal activity Content that expresses support or praise for groups, leaders, or individuals involved in these activities is enforced against. Also, the following people (whether living or deceased) and groups cannot maintain a presence (for example, have an account, Page or group) on Facebook: terrorist organisations, terrorists, hate organisations (and their leaders and prominent members) and mass and multiple murderers. One cannot have a Messenger account without a Facebook account. Terrorist organisations and terrorists include any non-state actor that: • Engages in, advocates or lends substantial support to purposive and planned acts of violence, • Which causes or attempts to cause death, injury or serious harm to civilians, or any other person not taking direct part in the hostilities in a situation of armed conflict, and/or significant damage to property linked to death, serious injury or serious harm to civilians • With the intent to coerce, intimidate and/or influence a civilian population, government or international organisation • In order to achieve a political, religious or ideological aim. A hate organisation is defined as any association of three or more people that is organised under a name, sign or symbol and that has an ideology, statements or physical actions that attack individuals based on characteristics, including race, religious affiliation, nationality, ethnicity, gender, sex, sexual orientation, serious disease or disability. A homicide is considered to be a mass murder if it results in three or more deaths in one incident. Any individual who has committed two or more murders over multiple incidents or locations is deemed a multiple murderer. Messenger prohibits in Messenger group profile pictures any symbols that represent any of the above organisations or individuals, unless they are shared with context that condemns or neutrally discusses the content. Messenger also takes action against users who it becomes aware of sharing content that:

	• praises any of the above organisations or individuals or any acts committed by them is prohibited
	• represents or supports in any way events that it designates as terrorist attacks, hate crimes, or mass shootings.
	• coordinates support for any of the above organisations or individuals or any acts committed by them we will take action against users.
	
	Under the section titled 'Violence and Incitement' of Facebook's Community Standards (Facebook, n.d.[2]), Messenger also takes action against users when Messenger is aware they are sharing language that incites or facilitates serious violence. In particular, users cannot share:
	
	• Threats that could lead to death (and other forms of high-severity violence) of any target(s), where threat is defined as any of the following:
	
	  • Statements of intent to commit high-severity violence
	  • Calls for high-severity violence including content where no target is specified but a symbol represents the target and/or includes a visual of an armament to represent violence; or
	  • Statements advocating for high-severity violence; or
	  • Aspirational or conditional statements to commit high-severity violence
	
	• Content that asks or offers services for hire to kill others (for example, hitmen, mercenaries, assassins) or advocates for the use of a hitman, mercenary or assassin against a target.
	• Admissions, statements of intent or advocacy, calls to action or aspirational or conditional statements to kidnap a target.
	• Threats that lead to serious injury (mid-severity violence) towards private individuals, minor public figures, vulnerable persons or vulnerable groups, where threat is defined as any of the following:
	
	  • Statements of intent to commit violence
	  • Statements advocating violence; or
	  • Calls for mid-severity violence including content where no target is specified but a symbol represents the target; or

	• Aspirational or conditional statements to commit violence; or • Content about other target(s) apart from private individuals, minor public figures, vulnerable persons or vulnerable groups and any credible: o Statements of intent to commit violence; o Calls for action of violence; o Statements advocating for violence; or o Aspirational or conditional statements to commit violence • Threats that lead to physical harm (or other forms of lower-severity violence) towards private individuals (self-reporting required) or minor public figures, where threat is defined as any of the following: • Statements of intent • calls for action • advocating, aspirational, or conditional statements to commit low-severity violence • Imagery of private individuals or minor public figures that has been manipulated to include threats of violence either in text or pictorially (adding bullseye, dart, gun to head etc.) • Any content created for the express purpose of outing an individual as a member of a designated and recognisable at-risk group • Instructions on how to make or use weapons if there is evidence of a goal to seriously injure or kill people, through: • Language explicitly stating that goal, or • photos or videos that show or simulate the end result (serious injury or death) as part of the instruction, • unless the aforementioned content is shared as part of recreational self-defence, for military training purposes, commercial video games or news coverage (posted by Page or with news logo) • Providing instructions on how to make or use explosives, unless there is clear context that the content is for a non-violent purpose (for example, part of commercial video games, clear scientific/educational purpose, fireworks or specifically for fishing)

	• Any content containing statements of intent, calls for action or advocating for high or mid-severity violence due to voting, voter registration or the outcome of an election • Misinformation that contributes to imminent violence or physical harm; and • Calls to action, statements of intent to bring armaments to locations, including but not limited to places of worship, or encouraging others to do the same.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.facebook.com/communitystandards/. also, for developers here: https://developers.facebook.com/devpolicy/ Terms of Service here: https://www.facebook.com/terms.php
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	See Section 4 of the Facebook Profile.
4.1 Notifications of removals or other enforcement decisions	See Section 4.1 of the Facebook Profile.
4.2 Appeal processes against removals or other enforcement decisions	See Section 4.2 of the Facebook Profile.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	See Section 5 of the Facebook Profile.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	See Section 6 of the Facebook Profile.

7. Does the service issue transparency reports (TRs) on TVEC	See Section 7 of the Facebook Profile.
8. What information/fields of data are included in the TRs?	See Section 8 of the Facebook Profile.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	See Section 9 of the Facebook Profile.
10. Frequency/timing with which TRs are issued	See Section 10 of the Facebook Profile.
11. Has this service been used to post TVEC?	Yes. See above sections 7-8 of the Facebook Profile.

5. iMessage/FaceTime

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There is no specific definition. However, Apple's Media Services Terms and Conditions (which govern iMessage and FaceTime) prohibit users from posting objectionable, offensive, unlawful, deceptive or harmful content, such as comments, pictures, videos, and podcasts (including associated metadata and artwork).
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.apple.com/ca/legal/internet-services/itunes/ca/terms.html
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	No procedures are specified. Apple broadly states that it may monitor and decide to remove or edit any submitted material.

4.1 Notifications of removals or other enforcement decisions	No notifications are specified.
4.2 Appeal processes against removals or other enforcement decisions	No appeal processes are specified.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Apple has a reporting mechanism that allow users to report content that violates its Submission Guidelines (included in Apple's Media Services Terms and Conditions). These reports are verified and processed by Apple's team. Given that iMessage and FaceTime are encrypted, it is difficult to see how an algorithm or an on-staff reviewer who works for Apple could detect any problematic content, including TVEC. The marginal economic costs of using human moderators to identify problematic content are probably relatively high. Apple is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	If Apple determines there is a breach or suspected breach of any of the provisions of its ToS, Apple may, without notice to the user, terminate the user's Apple ID, license to Apple's software and/or access to its services, which include iMessage and FaceTime.
7. Does the service issue transparency reports (TRs) on TVEC?	No. Apple does issue transparency reports (Apple, n.d.) that contain a section on content removal requests from governments and private parties reporting violations of its ToS or local laws, but there is no specific information on TVEC.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Possibly. A security manual issued by ISIS recommended use of iMessage to protect supporters' identities, (Zetter, 2015) but there is no evidence that ISIS supporters have actually used it (Dilger, 2015). Also, the FBI recently managed to unlock the iPhone of the perpetrator of the Pensacola attack, finding that he had been in contact with al-

	Qaeda 'using end-to-end encrypted apps.' However, it is not clear whether iMessage or FaceTime were actually used (Sky News, 2020).

6. WeChat

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There is no definition. However, in its Acceptable Use Policy, WeChat prohibits its users from submitting, uploading, transmitting or displaying any content which in fact or in WeChat's reasonable opinion: • breaches any laws or regulations (or may result in a breach of any laws or regulations); • creates a risk of loss or damage to any person; • harms or exploits any person (whether adult or minor) in any way, including via bullying, harassment or threats of violence; and • is hateful, harassing, abusive, racially or ethnically offensive, defamatory, humiliating to other people (publicly or otherwise), threatening, profane or otherwise objectionable.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.wechat.com/en/service_terms.html and https://www.wechat.com/en/acceptable_use_policy.html (Tencent, n.d.)
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	WeChat broadly states that it may review (but make no commitment to review) content (including any content posted by WeChat users) or third party programs or services made available through WeChat to determine whether or not they comply with WeChat's policies, applicable laws and regulations or are otherwise objectionable, and WeChat reserves the right to block or remove content for any reason, as required by applicable laws and regulations.
4.1 Notifications of removals or other enforcement decisions	No notifications are specified.

4.2 Appeal processes against removals or other enforcement decisions	No appeal processes are specified.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	WeChat provides no information in this regard. It has been reported that Chinese online firms, including WeChat, have a team of moderators policing problematic content. [47] Political activists have reported having been followed based on what they have said on WeChat, and chat records have turned up as evidence in court (Zhong, 2018). Also, research has shown that WeChat uses algorithmic technology (Knockel J. L.-N., 2018), keyword filtering and URL blocking (Ruan L. J.-N., 2016) to censor content that is in violation of its ToS (which may include the posting of TVEC). Although these methods had been reportedly applied only to accounts registered to mainland China phone numbers (Ruan L. J.-N., 2016), recent research has shown that international (i.e. non-Chinese) accounts are also monitored 'to invisibly train and build up WeChat's Chinese political censorship system' (Knockel, et al., 2020) The marginal economic costs of using automated tools to identify problematic content are probably very low (although fixed costs may be substantial), whereas the marginal economic costs of using human moderators to this end are probably relatively high. WeChat is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	WeChat notes that it may suspend or terminate access to WeChat if it reasonably believes that a user has breached WeChat's ToS, their use of WeChat creates risk for WeChat or other WeChat users, the suspension or termination is required by applicable laws, or at WeChat's sole and absolute discretion.
7. Does the service issue transparency reports (TRs) on TVEC	No.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.

10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Yes. The Christchurch shooting was posted on WeChat (Kenny, 2019). In addition, WeChat has been used to disseminate anti-Muslim propaganda (Huang, 2018).

7. Instagram

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	Instagram's Community Guidelines provide that Instagram is not a place to support or praise terrorism, organized crime, or hate groups, or to encourage violence or attack anyone based on their race, ethnicity, national origin, sex, gender, gender identity, sexual orientation, religious affiliation, disabilities, or diseases. Serious threats of harm to public and personal safety are also prohibited, as well as the sharing of graphic images to glorify violence. Though not identical, Instagram uses common implementation standards to interpret both Instagram's Community Guidelines and Facebook's Community Standards. To obtain more detail about Instagram's Community Guidelines, users can look to Facebook's Community Standards, which Instagram's Community Guidelines link to directly in several places. For example, Instagram's Community Guidelines state that "Instagram is not a place to support or praise terrorism, organized crime, or hate groups," and provide a direct link to Facebook's Community Standards' more detailed explanation of the policy rationale.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Instagram's Community Guidelines are available at https://help.instagram.com/477434105621119?helpref=page_content Instagram's ToS are available at https://help.instagram.com/581066165581870
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or	Instagram may remove content if it violates its Community Guidelines, or it may disable or terminate an account.

other enforcement decisions and appeal processes against them?	
4.1 Notifications of removals or other enforcement decisions	Instagram notifies the affected user of such content removals or account suspension or termination.
4.2 Appeal processes against removals or other enforcement decisions	If users believe their content has been removed or their account has been terminated in error, they can appeal the decision. It is possible for users to appeal the removal of content that was deemed to violate Instagram's 'counter-terrorism' policies (which are not specified). If content is found to have been removed in error, Instagram will restore the post and remove the violation from the account's record. In February 2020, Instagram rolled out a streamlined appeals process for disabled accounts directly through the app, instead of through the Instagram Help Center. See https://about.instagram.com/blog/announcements/safer-internet-day-2020/
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Instagram has implemented a built-in reporting option, so users may report content that violates the Community Guidelines. Instagram has a global team that reviews those reports and removes content that violates its guidelines. Instagram discloses that it may work with law enforcement, including when it believes that there is risk of physical harm or threat to public safety. The document titled 'Understanding the Community Standards Report' (Facebook) clarifies that Instagram uses the same methods as Facebook to identify and remove objectionable content, including TVEC. The marginal economic costs of using automated tools to identify TVEC are probably very low (although fixed costs may be substantial), whereas the marginal economic costs of using human moderators to this end are probably relatively high. Instagram is a member of the GIFCT and participates in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of ToS or Community Guidelines/Standards	Instagram can remove any content or information users share on the platform if Instagram believes that it violates its ToS and other policies (including the Instagram Community Guidelines). Instagram can also refuse to provide or can stop providing all or part of its service to a user (including terminating or disabling their account) immediately if the user clearly, seriously or repeatedly violates Instagram's ToS and other policies (including the Instagram Community Guidelines). Recently, Instagram announced an update of its account disabling policy, explaining that in addition to removing

	accounts with a certain percentage of violating content (which is undisclosed), it will also remove accounts with a certain number of violations within a window of time (also undisclosed) (Instagram, 2019).
7. Does the service issue transparency reports (TRs) on TVEC?	Yes. Facebook's last Community Standards Enforcement Report (Q2 2020) includes information from Instagram on the following areas: "adult nudity and sexual activity", "bullying and harassment", "child nudity and sexual exploitation," "regulated goods," "suicide and self-injury", "violent and graphic content", "hate speech" and "dangerous organisations: terrorism and organised hate".
8. What information/fields of data are included in the TRs?	The topic of "dangerous organisations: terrorism and organized hate" contains five fields of information: - *Prevalence (How prevalent were terrorism and violence and graphic content violations on Instagram?)* The prevalence metric is the percentage of views that included terrorism violations. For example, Instagram estimated an upper limit of 0.05% of views of content that violated its terrorism standards in Q2 2020. That means that out of every 10,000 views for the terrorism policy on Instagram, no more than 5 of those views contained content that violated that policy. (The figures refer to final determinations, not content that was initially flagged as a possible violation but may have been subsequently determined to be permissible.) - *Content actioned (How much content did Instagram take action on?)* Taking action may include removing a piece of content from Instagram, covering photos or videos that may be disturbing to some audiences with a warning, or disabling accounts. Content actioned is the total number of pieces of content that Instagram took action on during a given reporting period because it violated its community standards (in this case the terrorism policy). - *Proactive rate (Of the violating content actioned, how much did Instagram find before users reported it?)* This metric shows the percentage of content actioned for dangerous organisations content that Instagram found and flagged before users reported it. It counts detections made by both Facebook's AI tools and human reviewers. - *Appealed Content (How much of the content Instagram actioned did people appeal?)* This metric counts the number of pieces of content actioned for which people requested another review during the reporting period.

	- *Restored Content (How much content did Instagram restore after removing it?)* Restored content is the number of pieces of content that Instagram restored during the reporting period after previously actioning it. Instagram also includes recent trends regarding content actioned for organised hate and terrorism. For example, its last transparency report notes that content actioned for organized hate increased from 175.1K pieces of content in Q1 2020 to 266K in Q2 2020, and content actioned for terrorism decreased in Q2, from 440.6K pieces of content to 388.8K
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	- *Prevalence.* The prevalence metric is the estimated number of views of violating content, divided by the estimated number of total content views on Instagram, per reporting period. For example, if the prevalence of dangerous organisations is 0.18% to 0.20%, that means of every 10,000 content views, 18 to 20 on average were of content that violated Instagram's standards for dangerous organisations. The prevalence metric provides an indication of how often prohibited content is seen, rather than the total amount of such content published. For terrorism violations, in particular, Instagram only estimates the upper limit, which means that Instagram is 'confident that the prevalence of violating views is below that limit.' (Facebook). - *Content actioned.* Content actioned is the total number of pieces of content that Instagram took action on during a given reporting period because it violated its content policies. Instagram does not count those scenarios where it escalates content to law enforcement. This metric includes both content Instagram actioned after someone reported it and content that Instagram found proactively. - *Proactive rate.* This metric is calculated as: the number of pieces of content actioned that Instagram found and flagged before users reported them, divided by the total number of pieces of content actioned. - *Appealed Content.* This metric counts the number of pieces of content actioned for which people requested another review during the reporting period. Instagram observes that this metric shows the number of pieces of content that were appealed within the quarter, whereas restored content (see below) counts the content restored within the quarter. Because some appealed content may be restored in the following quarter, and some restored content was appealed in a previous quarter, these metrics cannot be directly compared.

	- *Restored content.* To arrive at this metric, Instagram counts the number of pieces of content that it restored during the reporting period after previously actioning it. Instagram may restore content either when a decision to remove is appealed or when Instagram discovers a reason to restore the content.
10. Frequency/timing with which TRs are issued	Instagram TRs are issued jointly with Facebook's and follow the same reporting schedule.
11. Has this service been used to post TVEC?	Yes. The media has covered many examples, (Carmen, 2015) (Hymas, 2019) (Cox, 2019).

8. TikTok

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There is no specific definition. However, TikTok's Community Guidelines provide that 'dangerous individuals or organisations' cannot use Tiktok to promote terrorism, crime, or other types of behaviour that could cause harm. Terrorists and terrorist organisations are expressly included within that group. TikTok defines 'terrorists and terrorist organisations' as any non-state actors that use premeditated violence or threats of violence to cause harm to non-combatant individuals, in order to intimidate or threaten a population, government, or international organisation in the pursuit of political, religious, ethnic, or ideological objectives. More broadly, TikTok defines 'dangerous individuals and organisations' as those that commit crimes or cause other types of severe harm. The types of groups and crimes include, but are not limited to Hate groups, Violent extremist organizations, Homicide, Human trafficking, Organ trafficking, Arms trafficking, Drug trafficking, Kidnapping, Extortion, Blackmailing, Money laundering, Fraud, Cybercrime. Names, symbols, logos, flags, slogans, uniforms, gestures, portraits, or other objects meant to represent dangerous individuals and/or organisations, or content that praises, glorifies, or supports dangerous individuals and/or organisations is prohibited on TikTok, except for educational, historical, satirical, artistic, and other content that can be clearly identified as counterspeech or aims to raise awareness of the harm caused by dangerous individuals and/or organisations.

	In addition, TikTok bans 'violent and graphic content', that is, content that is excessively gruesome or shocking, especially that promotes or glorifies abject violence or suffering. Some exceptions are allowed, for example, content that is newsworthy or meant to raise awareness about certain issues. Examples of content that is gratuitously shocking, sadistic or excessively graphic are depictions of violent or accidental deaths involving real people, depictions of dismembered, mutilated, charred, or burned human remains, depictions of gore in which an open wound or injury is the core focus, and depictions of severe physical violence. Similarly, content that attacks or incites violence against an individual or a group of individuals on the basis of protected attributes, including hate speech, is prohibited from TikTok. This includes content that verbally or physically threatens violence or depicts harm to an individual or a group based on any of the following protected attributes: race, ethnicity, national origin, religion, caste, sexual orientation, sex, gender, gender identity, serious disease or disability and immigration status. Also, TikTok prohibits content that dehumanizes or incites violence or hatred against individuals or groups, based on the foregoing attributes, including content claiming that they are physically or morally inferior. calling for or justifying violence against them, claiming that they are criminals. referring negatively to them as animals, inanimate objects, or other non-human entities, and promoting or justifying exclusion, segregation, or discrimination against them. Finally, TikTok prohibits content featuring 'hateful ideologies' (which are not defined), including content that promotes any hateful ideologies by talking positively about or displaying logos, symbols, flags, slogans, uniforms, salutes, gestures, portraits, illustrations, or names of individuals related to these ideologies; content that denies well-documented and violent events have taken place; and music or lyrics that promote hateful ideologies.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.tiktok.com/en/terms-of-use#terms-eea and https://www.tiktok.com/community-guidelines?lang=en
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.

4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	TikTok broadly states that it may, at any time and without prior notice, remove or disable access to content at its discretion for any reason or no reason. The removal of content may be based on TikTok finding the content objectionable, in violation of its ToS or Community Guidelines, or otherwise harmful to its services or users.
4.1 Notifications of removals or other enforcement decisions	No notifications are specified.
4.2 Appeal processes against removals or other enforcement decisions	If a user believes Tiktok has removed their content by mistake, they can appeal this decision.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	TikTok uses a combination of technology and content moderation to identify and remove content and accounts that violate its guidelines: Technology: TikTok has developed systems to automatically flag certain types of content that may violate its Community Guidelines. These systems take into account things like patterns or behavioural signals to flag potentially violative content, which allows TikTok to take swift action and reduce potential harm. TikTok notes that it regularly studies evolving trends, academic learnings, and industry best practices to continually enhance its systems. Content moderation: Technology today is not so advanced to be able to rely on it to enforce TikTok's policies. For instance, context can be important when determining whether certain content, like satire, is violative. As such, TikTok's team of trained moderators helps to review and remove content that violates TikTok's standards. In some cases, this team proactively removes evolving or trending violative content, such as dangerous challenges or harmful misinformation. Another way TikTok moderates content is based on reports receive from its users. TikTok's in-app reporting feature allows a user to choose from a list of reasons why they think something might violate TikTok's guidelines (such as violence or harm, harassment, or hate speech). If TikTok's moderators determine there's a violation, the content is removed. TikTok also works with a range of trusted experts to help it understand the dynamic policy landscape and develop policies and moderation strategies to address problematic content and behaviour as they emerge. These include the eight individual experts on TikTok's U.S. Content Advisory Council, and organisations such as ConnectSafely.org, the

	National Center for Missing and Exploited Children, WePROTECT Global Alliance, and others (TikTok, 2019-2020). The marginal economic costs of using automated tools to identify TVEC are probably very low (although fixed costs may be substantial), whereas the marginal economic costs of using human moderators to this end are probably relatively high. TikTok is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Violation of the Community Guidelines may result in account suspension, termination and/or content removal.
7. Does the service issue transparency reports (TRs) on TVEC?	Not specifically. However, in its second TR, TikTok informed that at the end of 2019 it started to roll out a new content moderation infrastructure that enables more transparency in the reporting of the reasons that videos are removed from TikTok. Under this infrastructure, when a video violates TikToks' Community Guidelines, it is labelled with the policy or policies it violates and is taken down. This means the same video may appear across multiple policy categories, including dangerous organisations, which includes terrorist and terrorist organisations (and by extension TVEC).

8. What information/fields of data are included in the TRs?	TikTok was only able to provide metrics for the month of December 2019, when its new content moderation infrastructure became effective. In particular, TikTok reported:
	
	- The percentage of videos taken down for violations under the following categories:
	  o adult nudity and sexual activities;
	  o minor safety;
	  o illegal activities and regulated goods,
	  o suicide, self-harm and dangerous acts
	  o violent and graphic content
	  o harassment and bullying
	  o hate speech, integrity and authenticity and dangerous individuals and organisations.
	
	- The number of videos removed globally for violating TikTok's Community Guidelines and/or ToS;
	
	- The percentage of videos that were proactively caught and removed by TikTok's systems before a user reported them; and

	- The percentage of videos taken down before receiving any views.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	No information is provided.
10. Frequency/timing with which TRs are issued	On a half-yearly basis.
11. Has this service been used to post TVEC?	Yes, see Sections 7-8 above.

9. QQ

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There is no definition. However, in its ToS, QQ prohibits its users from submitting, uploading, transmitting or displaying any content which in fact or in QQ's reasonable opinion: • breaches any laws or regulations (or may result in a breach of any laws or regulations); • creates a risk of loss or damage to any person; • harms or exploits any person (whether adult or minor) in any way, including via bullying, harassment or threats of violence; and • is hateful, harassing, abusive, racially or ethnically offensive, defamatory, humiliating to other people (publicly or otherwise), threatening, profane or otherwise objectionable.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.tencent.com/en-us/zc/termsofservice.shtml and https://www.tencent.com/en-us/zc/acceptableusepolicy.shtml[48]
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or	QQ broadly states that it may review (but make no commitment to review) content (including any content posted by users) or third party services made available through QQ to determine whether or not they comply with QQ's policies, applicable laws and regulations or are otherwise objectionable, and

other enforcement decisions and appeal processes against them?	QQ reserves the right to block or remove content for any reason, as required by applicable laws and regulations.
4.1 Notifications of removals or other enforcement decisions	No notifications are specified.
4.2 Appeal processes against removals or other enforcement decisions	No appeal processes are specified.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	QQ provides no information in this regard. QQ is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	QQ may suspend or terminate access to QQ if it reasonably believes that a user has breached QQ's ToS, their use of QQ creates risk for QQ or other QQ users, the suspension or termination is required by applicable laws, or at QQ's sole and absolute discretion.
7. Does the service issue transparency reports (TRs) on TVEC?	No.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Unknown.

10. Youku Tudou

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No definition is provided. However, in its ToS, Youku Tudou prohibits content that incites ethnic hatred, ethnic discrimination and/or undermines ethnic unity, as well as content that induces the commission of crimes, glorifies violence, or engages in terrorist activities.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at http://mapp.youku.com/service/agreement-eng
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	Youku Tudou broadly states that it 'manages' the information users upload, release or transmit on the platform, and takes measures such as suspending transmissions, removing uploaded content to prevent further dissemination, saving records and reporting to competent authorities in the event that information uploaded is banned by applicable laws and regulations or constitutes a breach of the ToS.
4.1 Notifications of removals or other enforcement decisions	No notifications are specified.
4.2 Appeal processes against removals or other enforcement decisions	No appeal processes are specified.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Youku Tudou provides no information in this regard. Youku Tudou is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Breaches of Youku Tudou's ToS may lead to the removal of content, the blocking of content and information, the suspension, termination or cancelation of a user account, or any other measures that may be taken in accordance with the applicable regulations.
7. Does the service issue transparency reports (TRs) on TVEC?	No.

8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Unknown.

11. Weibo

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There is no specific definition. However, Weibo's ToS prohibit users from uploading, displaying and transmitting any content that is offensive, abusive, intimidating, racially discriminatory, malicious, violent or otherwise illegal.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.weibo.com/signup/v5/protocol
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	Weibo broadly states that its operators have the right to review, supervise and process the behaviour and information of Weibo users, including but not limited to user information (account information, personal information, etc.), content data (location, text, pictures, audio, video, trademarks, patents, publications, etc.), and user behaviour (relationships, comments, private letters, participation topics, participation activities, marketing information, complaints, etc.).
4.1 Notifications of removals or other enforcement decisions	No notifications are specified.

4.2 Appeal processes against removals or other enforcement decisions	No appeal processes are specified.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Weibo has a reporting mechanism that allow users to report unlawful or objectionable content. These reports are verified and processed by moderators. The marginal economic costs of using human moderators to identify objectionable content are probably relatively high. Weibo is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Violation of the ToS entitles Weibo to discontinue or terminate the provision of its services.
7. Does the service issue transparency reports (TRs) on TVEC?	No.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Yes. The Christchurch shooting was posted on Weibo (Kenny, 2019).

12. QZone

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There is no definition. However, QQ International's ToS[49] prohibit users from publishing, delivering, transmitting or storing any content that contravenes the law or any content that is inappropriate, insulting, obscene and violent.

2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://imqq.com/html/FAQ_en/html/Miscellaneous_1.html [50]
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	No procedure is specified.
4.1 Notifications of removals or other enforcement decisions	No notifications are specified.
4.2 Appeal processes against removals or other enforcement decisions	No appeal processes are specified.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	QQ International provides no information in this regard. QQ International is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	QQ International states that breach of its ToS entitles them to interrupt the user licence, stop the provision of services, apply use restrictions, reclaim the user's QQ account, carry out legal investigations and other relevant measures, taking into consideration the severity of the user's conduct, without prior notice to the user.
7. Does the service issue transparency reports (TRs) on TVEC?	No.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating	Not applicable.

the information/data included in the TRs	
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Unknown.

13. iQIYI

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There is no definition. However, iQIYI's ToS prohibit the promotion of terrorism, extremism (not specifically violent extremism), hatred, ethnic discrimination and dissemination of violence.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.iqiyi.com/user/register/protocol.html
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	iQIYI broadly state that it reserves the right to cancel users' access to its products and services, or their ability to create, upload, publish and disseminate content, without prior notice.
4.1 Notifications of removals or other enforcement decisions	No notifications are specified.
4.2 Appeal processes against removals or other enforcement decisions	No appeal processes are specified.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	iQIYI provides no information in this regard. iQIYI is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.

6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	iQIYI notes that violations of its ToS give iQIYI the right to suspend or cancel the infringer's account, and report certain violations to the authorities, where appropriate.
7. Does the service issue transparency reports (TRs) on TVEC?	No.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Unknown.

14. Reddit

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There is no specific definition. However, Reddit's Content Policy prohibits content that encourages, glorifies, incites, or calls for violence or physical harm against an individual or a group of people. Also, according to its Transparency Report, Reddit removes terrorist content.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.redditinc.com/policies/user-agreement and https://www.redditinc.com/policies/content-policy It is important to note that Reddit employs a layered moderation system. While the Content Policy above governs all content on Reddit, the site itself consists of thousands of individual communities that are created and moderated by users themselves, on a volunteer basis. These moderators set their own community rules, unique to each specific community depending on its topic, in addition to the sitewide Content Policy. These rules are clearly marked in the sidebars of each individual community.

3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	Yes. Available at https://www.redditinc.com/policies/broadcasting-content-policy. In addition to the normal Content Policy, livestreamed content on Reddit is also subject to additional rules: **No NSFW Content** Broadcasts on Reddit may not include NSFW ("Not Safe for Work") content. As noted in the Content Policy, this means content that contains nudity, pornography or sexually suggestive content, or graphic violence, which a reasonable viewer may not want to be seen accessing in a public or formal setting such as a workplace. **No Illegal or Dangerous Behavior** Broadcasts may not contain activities that are illegal, or that pose unreasonable risk of bodily harm to the stream subject or bystanders. **No Quarantine-Eligible Content** Broadcasts on Reddit may not include content that would otherwise trigger a Quarantine. As noted in the Content Policy, this means content that average 'redditors' may find highly offensive or upsetting, or which promotes hoaxes.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	At the sitewide level, Reddit administrators (paid Reddit employees) have a variety of different methods to enforce their rules, including: • Asking the user nicely to 'knock it off' • Asking the user less nicely • Temporary or permanent suspension of accounts • Removal of privileges from, or adding restrictions to, accounts • Adding restrictions to Reddit communities, such as adding "Not safe for work" tags or quarantining (see below) • Removal of content • Banning of Reddit communities In addition to the enforcement steps that Reddit administrators may take at the sitewide level, volunteer user-moderators also have a number of enforcement methods that they use to enforce rules at the community-specific level. This may include banning the user from that community (either permanently or temporarily), or removing their posts from the community. These actions happen independently of Reddit administrators. Quarantining (Reddit Inc., n.d.) is a measure applied to communities (essentially, groups that share

	common interests) that average users may find offensive or upsetting, or that are dedicated to promoting hoaxes that warrant additional scrutiny. Its purpose is to prevent the quarantined community's content from being accidentally viewed by those who do not knowingly wish to do so, or viewed without appropriate context. Quarantined communities display a warning that requires users to explicitly opt-in to viewing the content. They generate no revenue, do not appear in non-subscription-based feeds (e.g. Popular), and are not included in search or recommendations. Reddit may also enforce a number of additional product restrictions that exist currently or as it may develop in the future (e.g. removing custom styling tools).
4.1 Notifications of removals or other enforcement decisions	Users are notified when administrators take enforcement actions. In the case of sitewide account suspensions, notice is given via a private message. A visual reminder will also appear on each page a user visits during the duration of the suspension and any time a forbidden action is attempted, such as posting or commenting. In the case of actions against an individual account, such as a sitewide suspension, the user receives a message notifying of the suspension and the reason. The suspended user will also see a banner directly in the user interface alerting to the suspension. If a user is suspended by volunteer moderators from an individual subreddit, notification will also come in the form of a private message. More information about account suspensions is available at https://www.reddithelp.com/hc/en-us/articles/360045734511-My-account-was-suspended-for-violating-Reddit-s-Content-Policy. In cases where individual pieces of content are removed, they will be "tombstoned," indicating to the public that content which was previously available has been removed. In cases where an entire subreddit is removed, a tombstone page will notify visitors of the removal, and the rule violated.
4.2 Appeal processes against removals or other enforcement decisions	Whether applied against an individual piece of content, an account, or an entire Subreddit, actions taken by Reddit in response to Content Policy violations may be appealed by a simple form, available at https://www.reddit.com/appeals. Appeals are evaluated by Reddit employees, and are either granted (resulting in the reinstatement of the content/account/Subreddit), or denied.

	There is also a separate appeals process for subreddits placed in quarantine. To be removed from quarantine, community moderators (see section 5 below) may file an appeal. The appeal should include a detailed account of changes to community moderation practices (appropriate changes may vary from community to community and could include techniques such as adding more moderators, creating new rules, employing more aggressive auto-moderation tools, adjusting community styling, etc.). The appeal should also offer evidence of sustained, consistent enforcement of these changes over a period of at least one month, demonstrating meaningful reform of the community. Reddit may, in its sole discretion, delete or remove content at any time and for any reason, including for a violation of its ToS or Content Policy, or if the content otherwise creates liability for them. Whether applied against an individual account or an entire community, actions taken by Reddit in response to Content Policy violations may be appealed. Reddit employees evaluate the appeals.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Reddit relies on a regime of volunteer user-moderators. Moderating a Reddit community is an unofficial, unpaid position. Community creators are automatically that community's first moderators, and they may appoint other users to be moderators to help them as well. Reddit reserves the right to revoke or limit a user's ability to moderate at any time and for any reason or no reason, including for a breach of its ToS. Moderators must follow the Moderator Guidelines (Reddit Inc., 2017), and when they receive reports related to their community, they must take action to moderate by removing content and/or escalating to Reddit administrators for review. Moderators may create and enforce rules for the communities they moderate, provided that such rules do not conflict with Reddit's ToS and other policies. Moderators can set up AutoModerator, which is a site-wide moderation tool assisting the moderation of communities. It enables moderators to carry out certain tasks automatically, such as replying to posts with helpful comments like pointing users to subreddit rules and removing or tagging posts by domain or keyword (Reddit Inc., n.d.). In addition, especially trained Reddit employees are in charge of enforcing Reddit's Content Policy at the sitewide level. Finally, individual Reddit users themselves also participate in flagging and ranking questionable

	content. Users may report content to either community moderators or Reddit employees. Each user may also downvote a piece of content. Sufficient numbers of downvotes result in the downranking or hiding of the content. Reddit has internal tools to hash and prevent re-upload of new pieces of terrorist content identified. Reddit also automatically blocks URLs from domains known to be controlled or operated by designated terrorist organizations. The marginal economic costs of using automated tools to identify objectionable content are probably very low (although fixed costs may be substantial), whereas the marginal economic costs of using human moderators to this end are probably relatively high. Reddit incurs no costs with regard to user moderators. Reddit is not a member of the GIFCT, but does participate in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	A violation of Reddit's ToS or Content Policy may lead to the removal of the violating content and/or temporary suspension or permanent termination of the infringer's account (depending on the severity of the incident), status as a moderator, or ability to access or use Reddit's services. Moderators must also follow the Moderator Guidelines, and failing to comply with them also has consequences, including, for example, loss of certain functionalities or moderator privileges. Finally, in the case of communities, if the community itself is not in compliance with Reddit's Content Policy or Moderator Guidelines, the community may be quarantined or banned, depending on the scale or seriousness of the violations.
7. Does the service issue transparency reports (TRs) on TVEC?	Yes. Reddit issues annual Transparency reports that include a section on content removals based on violation of individual community rules or Reddit's Content Policy, which includes the posting of violent content. Reddit reports designated terrorist content as a subset of these removals. In its most recent report (2020), Reddit specifically reported that out of the total pieces of violent content removed (26,986), there were 557 pieces of designated foreign terrorist organisation content (as designated by the US Department of State) (Reddit Inc., 2020) In its 2018 report (Reddit Inc., 2018), Reddit explained that the vast majority (around 2/3) of total content removals on Reddit are executed within individual subreddits (communities) by subreddit moderators. These removals are largely based on individual subreddit rules that are unique to each community and

	set by the moderators and communities themselves; therefore these removals are not necessarily indicative of Content Policy violations. While there may be overlap between enforcement of these rules and Reddit's Content Policy, moderator actions are entirely separate from removals done by Reddit administrators.
8. What information/fields of data are included in the TRs?	The report discloses the overall number and percentage of pieces of content removed by subreddit moderators and by Reddit administrators for violations of the Content Policy, as well as for content manipulation (spam and other inauthentic activity); the number and percentage of Content Policy violations removed by subreddit moderators and by Reddit administrators divided by categories of violations (Harassment, Minor sexualization, Violent content, Involuntary porn, Controlled goods, Private information, Impersonation and Ban evasion) as well as by type of content (image, video, text, livestream, crosspost); the number of accounts removed and suspended by Reddit administrators for violations of the Content policy or content manipulation (spam); the number of subreddit removals (due to Content Policy violations or lack of moderation); the number of quarantined Communities; the number of user reports Reddit received for potential policy violations, and the percentage of such reports that resulted in action taken by Reddit Administrators; and total number of appeals received by Reddit, broken down into appeals granted and denied. The report also discloses government and law enforcement requests for content removal or account information disclosure received by Reddit, broken down by country, and whether the requests were complied with or not. Other types of legal removal requests by private parties (eg lawyers/solicitors) are included as well, also broken down by country and compliance. The report additionally contains a detailed reporting of copyright removal requests and actions taken under the Digital Millennium Copyright Act (US).
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not disclosed.
10. Frequency/timing with which TRs are issued	On a yearly basis.
11. Has this service been used to post TVEC?	Yes. The footage of the Christchurch attack was made available in one of Reddit's communities. (Hatmaker, 2019) This led to Reddit administrators banning the entire community in question from the site. See also Section 7 above.

15. Kuaishou

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There is no specific definition. However, Kuaishou's ToS prohibit users from uploading, downloading, sending or transmitting information in violation of China's legal system, including content inciting hatred or ethnic discrimination, or spreading violence, homicide and terror.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.kuaishou.com/about/policy
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	Kuaishou states that it has the right to check and verify the content uploaded or published by users according to governmental requirements, as well as the right to deal with content in accordance with applicable laws and regulations.
4.1 Notifications of removals or other enforcement decisions	No notifications are specified.
4.2 Appeal processes against removals or other enforcement decisions	There is an appeal process in case an account has been banned in error. The instructions are available at https://www.kuaishou.com/help/feedback/2664?categoryId=hot
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Kuaishou has a reporting mechanism that allow users to report unlawful or objectionable content. These reports are verified and processed by moderators. The marginal economic costs of using human moderators to identify objectionable content are probably relatively high. Kuaishou is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.

6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Violation of the ToS entitles Kuaishou to restrict or prohibit use of Kuaishou and related services, close or deactivate the infringer's account, and contact the competent authorities, if applicable.
7. Does the service issue transparency reports (TRs) on TVEC?	No.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Unknown.

16. Telegram

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No definition is provided. However, Telegram's ToS prohibit the promotion of violence on publicly viewable Telegram channels. Notably, that prohibition does not apply to 'Secret Chats'.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://telegram.org/tos
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other	No procedures are disclosed. Telegram states that if they receive a court order that confirms a user is a terrorist suspect, they may disclose that user's IP address and phone number to the relevant authorities. Telegram also states

enforcement decisions and appeal processes against them?	that so far, this has never happened (Telegram, n.d.).
4.1 Notifications of removals or other enforcement decisions	No notifications are specified.
4.2 Appeal processes against removals or other enforcement decisions	No appeal processes are specified.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Telegram allows users to report content that violates its policies. Telegram also has a team that polices content on public channels. Since 2016, Telegram operates a channel called 'ISIS Watch', which highlights its efforts to delete public channels and bots that promote terrorist content. The channel claims Telegram has removed over 200,000 ISIS public channels and bots (Telegram, n.d.). The marginal economic costs of using human moderators to identify problematic content are probably relatively high. Telegram is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	No sanctions are specified.
7. Does the service issue transparency reports (TRs) on TVEC?	No.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Yes. Several terrorist attacks have been coordinated on Telegram (Bennett, 2019) (Hayden, Far-Right Extremists Are Calling for Terrorism on the Messaging App Telegram, 2019) (Bennett, 2019) (Hayden, Far-Right Extremists Are Calling

	for Terrorism on the Messaging App Telegram, 2019).

17. Snapchat

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No definition is provided. However, in Snapchat's Community Guidelines, under the heading 'Terrorism', Snap states that terrorist organisations are prohibited from using its platform, and Snapchat has no tolerance for content that advocates or advances terrorism. The term 'terrorist organisations' is not defined in Snap's public-facing guidelines, but internally, Snap applies this definition: "A **Foreign Terrorist Organization** is one designated as such by the U.S. State Department. **Terrorism** is the unlawful use of violence and intimidation, especially against civilians, in the pursuit of political aims." Snap also bans any content that promotes discrimination or violence on the basis of race, ethnicity, national origin, religion, sexual orientation, gender identity, disability or veteran status.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.snap.com/en-GB/terms/#terms-row and https://www.snap.com/en-GB/community-guidelines
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	Not applicable. Snapchat does not support livestreaming.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	Snap broadly states that it reserves the right to delete any content (i) which they think violates its ToS or Community Guidelines, or (ii) if doing so is necessary to comply with its legal obligations. Snap notes that they support the Santa Clara Principles on Transparency and Accountability in Content Moderation (Santa Clara University's High Tech Law Institute, n.d.), which state that companies should provide notice to users whose content is taken down or whose account is suspended about the reason for the removal or suspension. The Principles also state that companies should provide an opportunity for appeal of content removals and account suspensions, but there are as yet no content removal notifications and appeals against content

	removal decisions or account suspensions specified in Snapchat's policies.
4.1 Notifications of removals or other enforcement decisions	When the Trust & Safety team removes a Snap, the account holder receives a warning with a link to Snap's Community Guidelines. When this team takes action against an account, the account holder receives a notification that their account has been terminated for violation of Snap's Community Guidelines and/or Terms of Service.
4.2 Appeal processes against removals or other enforcement decisions	Depending on the infraction, a user may be banned from creating new accounts for six months or more. In order to appeal, a user may contact Snap via its support site.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Users are able to report content that violates Snapchat's policies (Snap Inc., n.d.). Platform Integrity evaluates potentially violating content based on user reports, trusted flaggers, internal escalations or automatic detection. Snap receives bulletins from the US National Counter Terrorism Center, alerts from Europol and law enforcement, and reports from third-party vendors concerning potential extremist content on Snapchat. Snap has a dedicated trust and safety team working on a 24/7 basis. Content that is found in violation of Snapchat's policies is removed. The marginal economic costs of using human moderators to detect objectionable content are probably relatively high. Snapchat is not a member of the GIFCT, but does participate in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Snaps that violate Snap's policies are removed. Similarly, accounts that repeatedly violate these policies and accounts that are dedicated to spreading violating material are removed from the platform. Users who violate Snap's policies repeatedly and/or egregiously may be banned from creating new accounts for six months or more. Generally, if a user violates Snapchat's ToS or Community Guidelines, Snapchat may remove the offending content, terminate the offender's account, and notify law enforcement. If a user's account is terminated for violations of Snapchat's policies, the infringer is prohibited from using Snapchat again.

7. Does the service issue transparency reports (TRs) on TVEC?	No. However, Snap does issue transparency reports (Snap Inc., 2015-2020) that contain a section on content removal requests from governments reporting violations of its ToS or Community Guidelines, which include a prohibition of TVEC. Yet, no TVEC-specific metrics are reported.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Snap issues transparency reports twice a year.
11. Has this service been used to post TVEC?	Yes. For example, footage of the terrorist attack in Nice, France in 2016 was disseminated on Snapchat's Live stories and Explorer features (Manileve, 2016).

18. Pinterest

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There is no specific definition. However, Pinterest's Community Guidelines provide that 'Dangerous organisations and individuals' are not allowed on Pinterest. These are groups that encourage, praise or provide aid to dangerous actors or groups and their activities, including: • Extremists, • Terrorist organisations, and • Gangs and other criminal organisations. The terms above are not defined. Also, Pinterest prohibits hateful content or the people and groups that promote hateful activities, including: • Slurs or negative stereotypes, caricatures and generalisations • Support for hate groups and people promoting hateful activities, prejudice and conspiracy theories • Condoning or trivialising violence because of a victim's membership in a vulnerable or protected group • Support for white supremacy, limiting

	women's rights and other discriminatory ideas • Hate-based conspiracy theories and misinformation, such as Holocaust denial • Denial of an individual's gender identity or sexual orientation, and support for conversion therapy and related programmes • Attacks on individuals including public figures based on their membership in a vulnerable or protected group • Mocking or attacking the beliefs, sacred symbols, movements or institutions of the protected or vulnerable groups identified below Protected and vulnerable groups include: people grouped together based on their actual or perceived race, colour, caste, ethnicity, immigration status, national origin, religion or faith, sex or gender identity, sexual orientation, disability, or medical condition. It also includes people who are grouped together based on lower socio-economic status, age, weight or size, pregnancy or ex-military status.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://policy.pinterest.com/en-gb/terms-of-service and https://policy.pinterest.com/en-gb/community-guidelines
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	Not applicable. Pinterest does not support live streamed content.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	Pinterest broadly states that it reserves the right to remove or modify user content, or change the way it is used in Pinterest, for any reason. This includes user content that is considered to be in violation of Pinterest's policies. Pinterest's Community Guidelines note that Pinterest limits the distribution of or remove content and accounts in cases of violations of its hateful activities and dangerous organisations and individuals policies.
4.1 Notifications of removals or other enforcement decisions	Pinterest notifies users when their content is removed 'in most cases', although it is not explained in which specific places notifications indeed take place.

4.2 Appeal processes against removals or other enforcement decisions	There are no appeal processes against a decision to remove content, but account suspensions can be appealed (Pinterest, n.d.).
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Pinterest has a reporting mechanism that allow users to report content that violates its policies. Pinterest has a team of moderators policing content. Terrorist and violent content is removed when detected. Pinterest informs that they collaborate with industry, government and security experts to identify terrorist groups. The marginal economic costs of using human moderators to detect objectionable content are probably relatively high. Pinterest is a member of the GIFCT, but does not participate in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	In case of violation of Pinterest's policies, Pinterest may terminate or suspend the violator's access to Pinterest immediately, without notice. Notifications of these actions take place at Pinterest's discretion.
7. Does the service issue transparency reports (TRs) on TVEC?	No. Pinterest does issue transparency reports (Pinterest, 2014-2020) that contain a section on content removal requests from governments and private parties reporting violations of its ToS or local laws, but there is no specific information on removals of TVEC.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Unknown.

19. Twitter

	There is no specific definition of terrorist or violent extremist *content*, but there is a specific policy on

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	Terrorism and Violent Extremism that includes information on what Twitter considers to be a terrorist or violent extremist organisation, along with examples of content that violates the company's Terrorism and Violent Extremism Policy. In the 'Safety' section of the 'Twitter Rules', terrorism and violent extremism are explicitly forbidden. Under Twitter's policy on Terrorism and Violent Extremism, users may not threaten or promote terrorism or violent extremism. Twitter asserts that there is no room in Twitter for terrorist organisations or violent extremist groups and individuals who affiliate with and promote their illicit activities. Twitter's assessments in this context are informed by national and international terrorism designations; however, these designations are not specified. Twitter also assesses organisations under its violent extremist group criteria. Organisations that: • identify through their stated purpose, publications, or actions as an extremist group; • have engaged in, or currently engage in, violence and/or the promotion of violence as a means to further their cause; and • target civilians in their acts and/or promotion of violence are deemed to be violent extremist groups. Twitter examines a group's activities both on and off Twitter to determine whether it engages in and/or promotes violence against civilians to advance a political, religious and/or social cause. Twitter provides the following examples of content that violates its Terrorism and Violent Extremism Policy: • engaging in or promoting acts on behalf of a terrorist organisation or violent extremist group; • recruiting for a terrorist organisation or violent extremist group; • providing or distributing services (e.g., financial, media/propaganda) to further a terrorist organisation's or violent extremist group's stated goals; and • using the insignia or symbols of terrorist organisations or violent extremist groups to promote them. In addition, Twitter's Hateful Conduct Policy provide that users may not promote violence against or directly attack or threaten other people on the basis

	of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. Accounts whose primary purpose is inciting harm towards others on the basis of these categories are prohibited. Also, users may not use hateful images or symbols in their profile image or profile header, nor may they use usernames, display names, or profile bios to engage in abusive behaviour, such as targeted harassment or expressing hate towards a person, group, or protected category. This policy bans violent threats, wishing, hoping or calling for serious harm on a person or group of people, references to mass murder, violent events, or specific means of violence where protected groups have been the primary targets or victims, inciting fear about a protected category, repeated and/or non-consensual slurs, epithets, racist and sexist tropes or other content that degrades someone, and hateful imagery (i.e. logos, symbols or images whose purpose is to promote hostility and malice against others based on their race, religion, disability, sexual orientation, gender identity or ethnicity/national origin). Lastly, Twitter's Glorification of Violence Policy prohibits the glorification of violence, especially violent events where people were targeted on the basis of their protected characteristics (including: race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease), as this could incite or lead to further violence motivated by hatred and intolerance. Under this policy, users cannot glorify, celebrate, praise or condone violent crimes, violent events where people were targeted because of their membership in a protected group, or the perpetrators of such acts. Glorification is defined to include praising, celebrating, or condoning statements, such as "I'm glad this happened", "This person is my hero", "I wish more people did things like this", or "I hope this inspires others to act". Violations of this policy include, but are not limited to, glorifying, praising, condoning, or celebrating: • violent acts committed by civilians that resulted in death or serious physical injury, e.g., murders, mass shootings; • attacks carried out by terrorist organizations or violent extremist groups; and • violent events that targeted protected groups, e.g., the Holocaust, Rwandan genocide.

2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://help.twitter.com/en/rules-and-policies/twitter-rules, https://help.twitter.com/en/rules-and-policies/violent-groups, https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy and https://help.twitter.com/en/rules-and-policies/glorification-of-violence
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	Twitter has a range of enforcement options that it may exercise when a user violates the Twitter Rules (Twitter, n.d.). a. *Tweet-level enforcement*: applies to content that violates Twitter's policies, but Twitter believes it is in the public interest that such content remain accessible. In this case, the tweet is hidden behind a notice that give users the option to view the content if they wish. These tweets of public interest are not available in the areas Top Tweets, safe search, recommendations via push and notifications tab, email and text recommendations, live event timeline and explore tab. Also, Twitter takes action at the Tweet level to ensure that it is not being overly harsh with an otherwise healthy account that made a mistake and violated its Rules. Possible tweet level measures include limiting tweet visibility, requiring tweet removal and hiding a violating tweet while awaiting its removal. b. *Direct message-level enforcement*: In a private direct message conversation, when a participant reports the other person, Twitter will stop the violator from sending messages to the person who reported them. The conversation will also be removed from the reporter's inbox. In a group direct message conversation, the violating direct message may be placed behind an interstitial to ensure no one else in the group can see it again.

	c. *Account-level enforcement*: applies when Twitter determines that a person has violated the Twitter Rules in a particularly egregious way, or has repeatedly violated them even after receiving notifications from Twitter. This may include: - Requiring media or profile edits: If an account's profile or media content is not compliant with Twitter's policies, Twitter may make it temporarily unavailable and require that the violator edit the media or information in their profile to come into compliance. Twitter also explains which policy their profile or media content has violated. - Placing an account in read-only mode: If it seems like an otherwise healthy account is in the middle of an abusive episode, Twitter might temporarily make their account read-only, limiting their ability to Tweet, Retweet, or Like content until calmer heads prevail. The person can read their timelines and will only be able to send Direct Messages to their followers. When an account is in read-only mode, others will still be able to see and engage with the account. The duration of this enforcement action can range from 12 hours to 7 days, depending on the nature of the violation. - Verifying account ownership: To ensure that violators do not abuse the anonymity Twitter offers and harass others on the platform, Twitter may require the account owner to verify ownership with a phone number or email address. This helps identify violators who are operating multiple accounts for abusive purposes and take action on such accounts. When an account has been locked pending completion of a challenge (such as being required to provide a phone number), it is removed from follower counts, Retweets, and likes until a phone number is provided. - Permanent suspension: This is the most severe enforcement action. Permanently suspending an

	account will remove it from global view, and the violator will not be allowed to create new accounts. When determining whether to take enforcement action, Twitter considers a number of factors, including (but not limited to) whether: • the behaviour is directed at an individual, group, or protected category of people; • the report has been filed by the target of the abuse or a bystander; • the user has a history of violating our policies; • the severity of the violation; • the content may be a topic of legitimate public interest (Twitter, n.d.).
4.1 Notifications of removals or other enforcement decisions	Notifications take place typically when Twitter requests a user to modify their behaviour and be in compliance with Twitter's rules (requiring media or profile edits), or in case of permanent account suspension. When Twitter permanently suspend an account, it notifies people that they have been suspended for abuse violations, and explains which policy or policies they have violated and which content was in violation.
4.2 Appeal processes against removals or other enforcement decisions	Users can appeal permanent suspensions if they believe Twitter made an error. Upon appeal, if it is found that a suspension is valid, Twitter responds to the appeal with information on the policy that the account has violated.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Twitter has three primary ways of detecting content that may violate its rules. 1. User reporting: Twitter encourages its users to report violations of the Twitter Rules. Moderators review the reports and decide whether the content in fact violates Twitter's rules. Twitter have a global team that manages enforcement of the Twitter Rules with 24/7 coverage in every supported language on Twitter. 2. Proactive content-based detections Twitter also uses internal, proprietary tools to detect violations of the Twitter Rules, including the posting of TVEC, based on the content that is being posted, for example known videos created by terrorist organisations. 3. Proactive behaviour-based detections Twitter utilises internal, proprietary tools to detect violations of the Twitter Rules, including the posting

	of TVEC, based on the behaviour exhibited that can be associated with terrorist organisations. Twitter has spoken of developing its anti-spam technology to proactively detect TVEC activity, given the tactics utilised by some groups is in part reminiscent of spam. The marginal economic costs of using automated tools to identify TVEC are probably very low (although fixed costs may be substantial), whereas the marginal economic costs of using human moderators to this end are probably relatively high. Twitter is member of the GIFCT and participates in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Violations of the Terrorism and Violent Extremism policy lead to the immediate and permanent suspension of the violating account. Violations of the Hateful Conduct Policy lead to different penalties, depending on a number of factors including, but not limited to, the severity of the violation and an individual's previous record of rule violations. For example, Twitter may ask someone to remove the violating content and serve a period of time in read-only mode before they can Tweet again. Subsequent violations will lead to longer read-only periods and may eventually result in permanent account suspension. If an account is engaging primarily in abusive behaviour, or is deemed to have shared a violent threat, Twitter will permanently suspend the account upon initial review. Violations of the Glorification of Violence Policy vary depending on the severity of the violation and the account's previous history of violations. The first time a user violates this policy, Twitter requires the user to remove the content. Twitter also temporarily locks the user out of his or her account. If a user continues to violate this policy after receiving a warning, the account will be permanently suspended.
7. Does the service issue transparency reports (TRs) on TVEC?	Yes. Twitter's Transparency Reports (Twitter, 2012-2020) include a section on Twitter Rules enforcement, which include the policies described in Section 1 above.
8. What information/fields of data are included in the TRs?	Twitter discloses the following metrics: • 'Accounts actioned': the number of unique accounts that were suspended or had some content removed for violating the Twitter Rules; • 'Content removed': the number of unique pieces of content (such as Tweets or an

	account's profile image, banner, or bio) that Twitter required account owners to remove for violating the Twitter Rules; and • 'Accounts suspended': the number of unique accounts that were suspended for violating the Twitter Rules. Each of the metrics above is broken down into the specific policies that comprise the Twitter rules, including those referenced in Section 1 (i.e. Terrorism and Violent Extremism, Hateful Conduct and Glorification of Violence). Specifically for Terrorism and Violent Extremism, Twitter reports the percentage of actioned accounts which were proactively identified and actioned. Twitter also includes trends in the reported data, some of which concern TVEC. For example, in its last report Twitter observed that there was a 9% decrease in the number of accounts actioned for violations of its Terrorism and Violent Extremism Policy as compared to the last reporting period.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	"Accounts Reported" reflects the total number of accounts that users reported as potentially violating the Twitter Rules. To provide meaningful metrics, Twitter de-duplicates accounts that were reported multiple times (whether multiple users reported an account for the same potential violation, or whether multiple users reported the same account for different potential violations). For the purposes of these metrics, Twitter similarly de-duplicates reports of specific Tweets. This means that even if Twitter receives reports about multiple Tweets by a single user, it counts these reports towards the "Accounts Reported" metric only once. "Accounts Actioned" reflects the total number of accounts that Twitter took some enforcement action on during the reporting period. Action may be any of the enforcement options explained in section 4 above. To provide meaningful metrics, Twitter de-duplicates accounts that were actioned multiple times for the same policy violation. This means that if Twitter took action on a Tweet or account under multiple policies, the account would be counted separately under each policy. However, if Twitter took action on a Tweet or account multiple times under the same policy (for example, Twitter may have placed an account in read-only mode temporarily and then later also required media or profile edits on the basis of the same violation), the account would be counted once under the relevant policy.

10. Frequency/timing with which TRs are issued	On a half-yearly basis.
11. Has this service been used to post TVEC?	Yes. See sections 7-8 above.

20. Douban

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There is no specific definition. However, Douban's ToS prohibit users from uploading, distributing and otherwise using content that contains gratuitous violence or promotes violence, racism, discrimination, bigotry, hatred or physical harm of any kind against any group or individual, or which is otherwise objectionable.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.douban.com/note/732773017/
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	Douban broadly states that it reserves the right (but have no obligation) to review any user content in its sole discretion. Douban also informs that it may remove or modify user content at any time for any reason, in its sole discretion, with or without notice to the relevant user.
4.1 Notifications of removals or other enforcement decisions	No notifications are specified.
4.2 Appeal processes against removals or other enforcement decisions	No appeal processes are specified.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	No information is provided. Douban is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.

6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Violations of the ToS entitle Douban to suspend the violator's rights to use its services or terminate the violator's account.
7. Does the service issue transparency reports (TRs) on TVEC?	No.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Unknown.

17. LinkedIn

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	LinkedIn's Professional Community Policy has the following sections that prohibit TVEC: **Do not threaten, incite, or promote violence:** We don't allow threatening or inciting violence of any kind. We don't allow individuals or groups that engage in or promote violence, property damage, or organized criminal activity. You may not use LinkedIn to express support for such individuals or groups or to otherwise glorify violence. **Do not share harmful or shocking material:** We don't allow content that is excessively gruesome or shocking. This includes content that is sadistic or gratuitously graphic, such as the depiction of severe physical violence. We don't allow content or activities that promote, organize, depict, or facilitate criminal activity. We also don't allow content depicting or promoting instructional weapon making, drug abuse, and threats of theft. Do not engage in or promote non-consensual sexually explicit content (e.g., revenge porn), escort services, prostitution, exploitation of children, or human trafficking. Do not share content or activities that promote or encourage suicide or any type of self-injury, including self-mutilation and eating disorders. If you see signs that someone may be considering self-harm, please **report** it. **Do not post terrorist content or promote terrorism:** We don't allow any terrorist organizations or violent extremist groups on our

	platform. And we don't allow any individuals who affiliate with such organizations or groups to promote their activities. Content that depicts terrorist activity, that is intended to recruit for terrorist organizations, or threatens, promotes, or supports terrorism in any manner is not tolerated.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.linkedin.com/legal/professional-community-policies
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	Yes. In addition to having to comply with the ToS and the LinkedIn Professional Community Policies, live streaming is a limited feature on LinkedIn. Any member who wants to use it must submit an application and be reviewed under a specific set of criteria. The application form is available here: https://www.linkedin.com/help/linkedin/ask/lv-app LinkedIn has provided additional best practices and guidelines for live streaming, which are available here: https://www.linkedin.com/help/linkedin/answer/100225?query=linkedin%20live&hcppcid=search
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	LinkedIn encourages users to report content that violates its Professional Community Policy. When a user reports another member's content, that other member is not told who made the report, and the reporting user no longer sees the content or conversation they reported in their feed or messaging inbox. LinkedIn may review the reported content or conversation to take additional measures like removing the content, or in the case of severe or repeated violations, suspending the author if the content is in violation of its ToS or policies. LinkedIn has build features to provide enhanced transparency to both reporters and authors when it makes content moderation decisions. One latest feature includes a feedback loop, which means reporters will get notification at the time of report as well as when LinkedIn makes a decision on the report, and authors whose content gets removed for policy violations will be notified at the time of removal and be provided with the ability to appeal. This feature is being rolled out first in the US, France and Canada and will later be expanded to the rest of the world. (See more here: https://blog.linkedin.com/2020/september/29/new-features-help-keep-it-professional)
4.1 Notifications of removals or other enforcement decisions	Content removals are notified to both the author of the content and the reporter of the content.

4.2 Appeal processes against removals or other enforcement decisions	If an account has been restricted or contented removed and the user believe the action was in error, the user can appeal the decision. Also, when an author receives a notification about their content being removed, the author can appeal the decision.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Users are able to report content that violates LinkedIn's policies. Moderators review the reports to decide whether to take further actions. LinkedIn's parent company, Microsoft, Inc., states that whenever terrorist content on its hosted consumer services is brought to its attention via its online reporting tool, it removes it (Microsoft, 2016). The marginal economic costs of using human moderators to detect objectionable content are probably relatively high. LinkedIn is a member of the GIFCT. It leverages the GIFCT hash-sharing database, as well as employs other machine classifiers and processers to detect potential TVEC on its platform.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	The posting of content that violates LinkedIn's ToS or other policies may lead to content removal, or in the case of severe or repeated violations, suspension of the author's account.
7. Does the service issue transparency reports (TRs) on TVEC?	Not specifically. LinkedIn issues bi-annual transparency reports (LinkedIn, n.d.[106]) that contain a section on content removal requests from governments reporting violations of its ToS or local laws, as well as a report on content removal under its Professional Community Policies. TVEC is reported as part of the "violent or graphic" category, which "includes content that threatens or promotes terrorism, violence, or other criminal activity, and content that is extremely violent or intended to shock or humiliate others" and thus is broader than TVEC alone. The latest report is available here: https://about.linkedin.com/transparency/community-report
8. What information/fields of data are included in the TRs?	Total content removed, as well as the specific number of content removed as "violent or graphic", which includes TVEC. LinkedIn also reports the total number of content removal requests from governments reporting violations of its ToS or local laws, by country, as well as the percentage of requests on which LinkedIn took action, but there is no specific information on removals of TVEC.
9. Methodologies for determining/calculating /estimating the	Broad explanations are provided in the Community Report: https://about.linkedin.com/transparency/community-report

information/data included in the TRs	
10. Frequency/timing with which TRs are issued	Every six months.
11. Has this service been used to post TVEC?	Possibly. Research has shown that U.S.-based extremists – though not necessarily violent extremists – have used LinkedIn to promote their agendas (START (National Consortium for the Study of Terrorism and Responses to Terrorism), 2018).

22. Baidu Tieba

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There is no specific definition. However, Baidu Tieba's ToS prohibits content that incites ethnic hatred and ethnic discrimination, as well as content that spreads violence, murder and terrorism.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://gsp0.baidu.com/5aAHeD3nKhI2p27j8IqW0jdnxx1xbK/tb/eula.html
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	No procedures are specified.
4.1 Notifications of removals or other enforcement decisions	No notifications are specified.

4.2 Appeal processes against removals or other enforcement decisions	No appeal processes are specified.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Baidu Tieba has a reporting mechanism that allow users to report unlawful or objectionable content. These reports are verified and processed by moderators, who ultimately make the decision to keep or remove the content. The marginal economic costs of using human moderators to detect objectionable content are probably relatively high. Baidu Tieba is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	If it deems that a user has violated its ToS, Baidu Tieba may apply a temporary or permanent ban on the infringer, suspend or delete the infringer's account, or impose any other penalties in accordance with applicable regulations.
7. Does the service issue transparency reports (TRs) on TVEC?	No.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Unknown.

23. Skype

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	Skype's parent company is Microsoft. Microsoft's Services Agreement, which governs Skype, prohibits any activity that is harmful to others, such as posting terrorist or violent extremist content, communicating hate speech or advocating violence against others. Microsoft has stated (Microsoft, 2016) that, for the purposes of its services, terrorist content is material posted by or in support of organizations included on

	the Consolidated United Nations Security Council Sanctions List (United Nations Security Council) that depicts graphic violence, encourages violent action, endorses a terrorist organization or its acts, or encourages people to join such groups. The U.N. Sanctions List includes a list of groups that the U.N. Security Council considers to be terrorist organizations. No definition of violent extremism is provided, but Skype's ToS prohibit users from submitting or publishing any content that is hateful, abusive, illegal, racist, offensive or otherwise objectionable in any way. In its Digital Safety Content Report (Microsoft, 2021), Microsoft clarifies that 'both terrorist and violent extremist content is prohibited on Microsoft platforms and services', and that Microsoft Services Agreement Code of Conduct prohibits the 'posting of terrorist or violent extremist content.'
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Microsoft's Services Agreement is available at : https://www.microsoft.com/en-us/servicesagreement See also https://www.skype.com/en/legal/ios/tos/#1
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	Skype specifies a notice and take-down procedure. If Skype receives a notification that any material a user posts, uploads, edits, hosts, shares and/or publishes on Skype (excluding private communications) is inappropriate, infringes any rights of any third party, or if Skype wishes to remove that material or content for any reason whatsoever, Skype reserves the right to automatically remove it for any reason immediately or within such other timescales as may be decided from time to time by Skype in its sole discretion. As described in Microsoft's Services Agreement, "If you violate these Terms, we may stop providing Services to you or we may close your Microsoft account. We may also block delivery of a communication (like email, file sharing or instant message) to or from the Services in an effort to enforce these Terms or we may remove or refuse to publish Your Content for any reason. When investigating alleged violations of these Terms, Microsoft reserves the right to review Your Content in order to resolve the issue."

4.1 Notifications of removals or other enforcement decisions	Notifications are at Microsoft's discretion. Microsoft's Services Agreement states: "When there's something we need to tell you about a Service you use, we'll send you Service notifications. If you gave us your email address or phone number in connection with your Microsoft account, then we may send Service notifications to you via email or via SMS (text message), including to verify your identity before registering your mobile phone number and verifying your purchases. We may also send you Service notifications by other means (for example by in-product messages)."
4.2 Appeal processes against removals or other enforcement decisions	Microsoft's Account suspension appeals form is available at: https://www.microsoft.com/en-us/concern/AccountReinstatement
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Microsoft states that the Microsoft Services Agreement Code of Conduct prohibits the "posting [of] terrorist or violent extremist content." Microsoft encourages the reporting of content posted by – or in support of – a terrorist organization that depicts graphic violence, encourages violent action, endorses a terrorist organization or its acts, or encourages people to join such groups. Microsoft reviews these reports; takes action on content; and, if necessary, suspends accounts associated with violations of our Code of Conduct. In addition, Microsoft leverages a variety of tools, including hash-matching technology and other forms of proactive detection, to detect terrorist and violent extremist content. When users file reports, moderators review them to decide whether further action is warranted. Microsoft states that whenever terrorist content on its hosted consumer services is brought to its attention via its online reporting tool, it removes it (Microsoft, 2016). Microsoft uses scanning technologies (e.g., PhotoDNA or MD5) and other AI-based technologies, such as text-based classifiers, image classifiers, and the grooming detection technique to detect TVEC (Microsoft, 2021) The marginal economic costs of using human moderators to detect objectionable content are probably relatively high. Microsoft is a founding member of the GIFCT and participates in GIFCT's Hash Sharing Consortium.

6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Posting content in violation of Skype's ToS or other policies may lead to the termination or suspension of the infringer's Skype account and use of Skype. See also information in Sections 4 and 4.1 above.
7. Does the service issue transparency reports (TRs) on TVEC	Yes. TVEC numbers for Skype are included in Microsoft's Digital Safety Content Report (Microsoft, 2021). This report is inclusive of Microsoft consumer products and services including (but not limited to) OneDrive, Outlook, Skype, Bing and Xbox. It must be noted that TVEC metrics are reported on aggregate for all Microsoft consumer services and products, and not on a per-product basis.
8. What information/fields of data are included in the TRs?	• Pieces of TVEC actioned • Number of accounts suspended due to TVEC • % of TVEC actioned that Microsoft detected • % of accounts suspended for TVEC that were reinstated upon appeal
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	"Content actioned" refers to when Microsoft removes a piece of user-generated content from its products and services and/or blocks user access to a piece of user-generated content. "Account suspension" means removing the user's ability to access the service account either permanently or temporarily "Proactive detection" refers to Microsoft-initiated flagging of content on its products or services, whether through automated or manual review.
10. Frequency/timing with which TRs are issued	Not reported.
11. Has this service been used to post TVEC?	Possibly. Research by the Counter Extremism Project has found that a number of individuals have accessed and disseminated official extremist (though the source does not expressly specify violent extremist) propaganda materials on Skype (Counter Terrorism Project).

24. Quora

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No definition is provided. However, in Quora's Be Nice, Be Respectful Policy, under the heading 'No glorifying or advocating violence', Quora states that it will ban and delete all the content of any user who is a confirmed and/or declared member of any group on the U.S. State Department list of Foreign Terrorist Organisations, or is a confirmed participant in acts of mass violence or hate crimes.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.quora.com/about/tos, https://www.quora.com/about/acceptable_use and https://www.quora.com/What-is-Quoras-Be-Nice-Be-Respectful-policy/answer/Quora-Official-Account
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	Quora states that it has the right but not the obligation to refuse to distribute any content on the Quora platform or to remove content. Violations of Quora's policies may lead to a content warning, and if the violator persists with their conduct, they may be prevented from asking questions, writing answers and making comments (edit-blocked) or they may be banned. (Quora, n.d.) Edit-blocks and bans may be temporary; if a person is banned or edit-blocked, they can come back when they cool off and decide to stop their behaviour. Edit-blocks generally last until the person responds via PM and makes their case to be unblocked.
4.1 Notifications of removals or other enforcement decisions	There are no notifications of content removal, but there are content warnings, as specified above.
4.2 Appeal processes against removals or other enforcement decisions	If a user feels that an edit-block or ban was imposed unfairly, then he or she can appeal Quora's decision.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Users are able to report content that they believe violates Quora's policies. Reports are sent to the Quora Moderation team for review. The marginal economic costs of using human moderators to detect objectionable content are probably relatively high. Quora is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.

6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Content that violates the Be Nice, Be Respectful policy may be reported to and removed by administrators, and violations of this policy can result in a warning, comment-blocking, an edit-block, or a ban (see section 4 above). Depending on the severity of the Be Nice, Be Respectful violation, a user may be banned immediately (i.e., without waiting for content warnings or edit-blocks). Also, Quora may terminate or suspend a user's Quora account for violating any Quora policy.
7. Does the service issue transparency reports (TRs) on TVEC?	No.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Yes. Questions about how to join a terrorist organisation have been posted on Quora (Lange, 2017).

25. Xigua

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There is no specific definition. However, Xigua's ToS prohibit users from promoting terrorism and extremism.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.ixigua.com/user_agreement/
3. Are there specific provisions applicable to livestreamed content in	No.

the ToS or Community Guidelines/Standards?	
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	No procedures are specified.
4.1 Notifications of removals or other enforcement decisions	No notifications are specified.
4.2 Appeal processes against removals or other enforcement decisions	No appeal processes are specified.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Users can report any type of unlawful activity or content on Xigua. Xigua is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Violation of Xigua's ToS may lead to the termination of the infringer's account and access to Xigua's services, without prior notice.
7. Does the service issue transparency reports (TRs) on TVEC?	No.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Unknown.

26. Viber

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There is no specific definition. However, Viber's Public Content Policy provides that overly graphic expressions of violence, in particular where the violence is glorified or encouraged, are not allowed on Viber. This includes extreme depictions or descriptions of violence and credible threats of violence to any individual and/or group. Viber prohibits planning or promoting violent acts that could directly or indirectly cause physical or mental harm to others.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.viber.com/terms/viber-terms-use/ and https://www.viber.com/terms/viber-public-content-policy/
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	Not applicable. Viber does not have a livestreaming feature currently.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	Viber states that upon creating a Community, the user automatically becomes a "Superadmin" of that Community. Administrators must ensure that all content uploaded and displayed in their Public Account or Community complies with Viber's policies, terms of service and all applicable laws and regulations. Administrators may not engage in or permit third parties to engage in any behaviour that is prohibited under any of them. Administrators have the ability to delete content themselves. Viber may remove any or all content if they deem that such content is unauthorized or illegal or violates Viber's Policies.
4.1 Notifications of removals or other enforcement decisions	No notifications are specified.
4.2 Appeal processes against removals or other enforcement decisions	Users may contact Viber's support to appeal a removal of content or blocking and Viber considers each request.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers,	Users have the option to report content that violates Viber's Content Policy. Viber reviews those reports and operates a moderation team to determine the most suitable course of action. Viber also has internal algorithms applied to detect certain illegal content.

hash-sharing/URL sharing database)	Administrators have the ability to remove violating content from their Accounts and Communities. It is difficult to determine the extent to which Viber is moderated. Viber's Terms of Use provide that Viber does not undertake to monitor Public Chats or other Forums, and assumes no liability for the content posted therein. In addition, Viber's core features are encrypted), for which reason moderation of content disseminated through those features is not possible. However, the public features such as communities and public chats are not end to end encrypted, and Viber can, upon reports, review them and if required remove them. The marginal economic costs of using human moderators to detect objectionable content are probably relatively high. User moderators entail no cost for Viber. Viber is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Content that violates Viber's policies or that Viber otherwise finds objectionable is removed. In those cases, Viber may suspend or terminate users' accounts, and block participants or block communities.
7. Does the service issue transparency reports (TRs) on TVEC?	No.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating /estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Yes. ISIS announced (Site Intelligence Group Enterprise, 2018) a Nashir News Agency (the ISIS-linked media dissemination group) account on Viber (Katz, A Growing Frontier for Terrorist Groups: Unsuspecting Chat Apps, 2019). Viber closed the account immediately after finding it.

27. Discord

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No definition is provided. However, Discord's Community Guidelines ban attacks on a person or a community based on attributes such as their race, ethnicity, national origin, sex, gender, sexual orientation, religious affiliation, or disabilities. Also, threats of violence or harm to others are prohibited. The use of Discord for the organization, promotion or support of violent extremism is also prohibited. Violent extremist content is defined in Discord's last transparency report as 'content where users advocate or support violence as a means to an ideological end.' Examples include racially motivated violent groups, religiously motivated groups dedicated to violence, and incel groups. In addition, Discord's ToS provide that users cannot defame, libel, ridicule, mock, stalk, threaten, harass, intimidate or abuse anyone.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://discordapp.com/terms and https://discordapp.com/guidelines
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	Discord explains violation of its Community Guidelines or other policies enables them to take a 'number of steps', which are specified in Section 6 below.
4.1 Notifications of removals or other enforcement decisions	No notifications are specified.
4.2 Appeal processes against removals or other enforcement decisions	Users are able to appeal actions taken against their accounts.
5. Means of identifying TVEC (for example, monitoring algorithms, user	Users can report any content that violates Discord's ToS and Guidelines. Discord has stated that, although it does not read users' private messages, it does investigate and take immediate appropriate

generated, human (staff) reviewers, hash-sharing/URL sharing database)	action against any reported ToS violation by a server (something akin to a group or community under a common theme) or user (Liao, 2018). After the report, Discord's 'Trust and Safety' team acts as detectives, looking through the available evidence and gathering as much information as possible. This investigation centres on the reported messages, but can expand if the evidence shows that there is a bigger violation — for example, if the entire server is dedicated to bad behaviour, or if the behaviour appears to extend historically. Discord uses "smart computers" and automation to detect spamming and exploitative content such as revenge porn, deep fakes and content threatening child safety, and implements systems such as PhotoDNA to detect that content. Discord's last transparency report suggests that these tools are also used to detect violent extremism (Discord, 2020). Discord has received reports of servers (something similar to groups of users gathered under a theme) focused on spreading hate speech, harassing others, and convincing others to follow dangerous ideologies. Discord states that they take these reports seriously and remove servers exhibiting extremist (not specifically violent extremist) behaviour. In addition, Discord asserts that it works with law enforcement agencies, third-parties (such as news outlets and academics), and organisations focused on fighting hate (like the Anti-Defamation League and Southern Poverty Law Center) to make sure Discord is up-to-date and ahead of any potential risks. The marginal economic costs of using automated tools to identify objectionable content are probably very low (although fixed costs may be substantial), whereas the marginal economic costs of using human moderators to this end are probably relatively high. Discord is a member of the GIFCT, but does not participate in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	If a violation of Discord's Community Guidelines is detected, Discord may take any of the following actions regarding users and/or servers: - Removing the content - Warning users and educating them about their violation - Temporary banning as a "cool-down" period

	- Permanently banning users from Discord and making it difficult for them to create another account - Removing a server from Discord - Disabling a server's ability to invite new users
7. Does the service issue transparency reports (TRs) on TVEC?	Yes, but to a very limited extent. Discord issued its first transparency report of any kind in 2019, (Discord, 2019) in which it disclosed the number of reports they received for violations of its Community Guidelines, which might have included the posting of TVEC, although this was not mentioned explicitly. Discord's second transparency report (covering the period April-December 2019) follows a similar structure as the first one, and contains some information on violent extremist content removal (Discord, 2020).
8. What information/fields of data are included in the TRs?	Discord's second transparency report discloses: - the overall number of reports received, as well at the percentage that fell within each prohibited category (e.g. self-harm, harassment, threats, and spam). It is not clear under which category violent extremism falls. - the percentage of the reports on which Discord took action, but does not disclose whether that action was content removal, a warning, or account deletion. - The total number of account bans and server bans, broken down by prohibited categories - The number of violent extremism server deletions by month, proactively detected with automated tools. - The number of accounts reinstated on appeal, broken down by prohibited categories.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	No information available.
10. Frequency/timing with which TRs are issued	Undefined. However, Discord informed in its last transparency report that it aims to follow a semi-annual publication schedule.

11. Has this service been used to post TVEC?	Yes. See Section 8 above.

28. Vimeo

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There is no definition. However, Vimeo prohibits any content that promotes or supports "terror or hate groups"; depicts unlawful acts or extreme violence; and provides instructions on how to assemble explosive/incendiary devices or homemade/improvised firearms. Furthermore, members of a "terror or hate group" cannot create a Vimeo account. The term "terror or hate groups" is not defined. Also, content violates Vimeo's anti-hate and anti-discrimination policy when it (1) is directed to a group based upon personal characteristics, such as race, religion, gender, and sexual orientation; (2) sends a message of inferiority; and (3) would be considered extremely offensive to a reasonable person. Vimeo's definition covers, for example, videos that assert harmful stereotypes, claim racial superiority of one group over another, or suggest that certain groups of people of a particular religion are involved in far-flung conspiracies (Cheah, 2019).
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://vimeo.com/terms and https://vimeo.com/help/guidelines
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	Vimeo states that context is of the essence in the application of its rules and processes. When prohibited content appears in the context of a news story or a narrative device in a dramatic work, Vimeo is likely to leave it up. If, however, the overall driving message of the work is to perpetuate a viewpoint that Vimeo has specifically banned, they will remove it. Vimeo also considers a user's speech outside Vimeo (such as social media platforms, blogs, or anywhere else their personal views are clearly represented) in making calls about intent and good faith (Cheah, 2019). As a rule, Vimeo moderators will remove videos that show people being murdered, tortured, or physically

	or sexually abused, or display shocking, disgusting, or gruesome images. That said, Vimeo understands that there can be videos that engage with these subjects in a critical, thoughtful way. Videos that report on real-world situations sometimes necessarily contain some graphic or violent scenes. Context is important, and documentary or journalistic videos have greater leeway when it comes to depicting violence or the aftermath of violence. To avoid being removed, videos with these elements may not be sensationalistic, exploitative, or gratuitous. They must also be marked with a "Mature" content rating. Videos that recruit for or propagandise terrorist organisations, regardless of whether they show actual violence, are never allowed (Vimeo , n.d.).
4.1 Notifications of removals or other enforcement decisions	Some content removal decisions are notified, such as removals due to copyright infringement. However, Vimeo does not provide users with notice of video or account removals (or a mechanism for appeal) when the removal involves certain categories of prohibited content, such as suspected child abuse material and terrorist content.
4.2 Appeal processes against removals or other enforcement decisions	Copyright-based removals may be appealed. However, there are no appeal processes against a decision to remove TVEC.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Users can report any content that violates Vimeo's guidelines and policies. Vimeo states that it may monitor users' accounts, content, and conduct, regardless of their privacy settings. Vimeo has signed an agreement with Active Fence to help identify TVEC content and expects to implement this partnership in early 2020. The marginal economic costs of using human moderators to detect objectionable content are probably relatively high. Vimeo is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	In case of violation of Vimeo's policies and ToS, Vimeo may, at its option, suspend, delete, or limit access to the infringer's account or any content within it; and terminate the infringing account.

7. Does the service issue transparency reports (TRs) on TVEC?	No.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Unknown.

29. IMO

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There is no definition. However, IMO's Acceptable Use Policy prohibit the use of its services to disseminate any threats of violence.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://imo.im/policies/terms_of_service and https://imo.im/policies/acceptable_use_policy.html
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	IMO broadly states that it reserves the right to remove, screen, edit, or disable access to any content, without notice to the user owning the content, that IMO considers in its sole discretion to be in violation of its policies or otherwise harmful to the IMO Service.
4.1 Notifications of removals or other enforcement decisions	No notifications are specified.
4.2 Appeal processes against removals or other enforcement decisions	No appeal processes are specified.

5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	IMO states that they are 'under no obligation to review' content, but it reserves the right to do so at any time. However, it is unclear what manner(s) of review they would undertake. IMO is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Violation of IMO's policies may result in the suspension or termination of the infringer's account.
7. Does the service issue transparency reports (TRs) on TVEC?	No.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Unknown.

30. LINE

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No definition is provided. However, LINE's ToS prohibit the posting or transmission of violent content. Also, 'activities that benefit or collaborate with anti-social groups' are not allowed. The term 'anti-social group' is not defined.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://terms.line.me/line_terms/
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	Yes, available at https://terms2.line.me/LINELIVE_ToC_ME1

4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	LINE discloses a two-step process to monitor posts on its Timeline, LINE LIVE, LINE Manga, LINE Fortune, LINE Pasha, LINE Step, LINE BLOG, LINE Delima and WizBall: First, user-posted content on supported LINE services is checked by LINE's automatic monitoring system to ensure that it does not contain any prohibited language, break any service rules, or violate LINE's ToS or any relevant laws. If objectionable content is found by the monitoring system, it is immediately suspended after being posted. Next, a monitoring team checks any content the monitoring system cannot classify. The monitoring team compares the content against a set of evaluation criteria and previous examples to make a decision on whether or not the content is permitted. If the monitoring team determines the posted content is in violation of LINE's ToS or any applicable laws, it is suspended (LINE, 2019-2020). LINE is unable to monitor any message a user sends/receives on a regular LINE chat room unless the user sends unencrypted chat data to LINE by using the reporting tool (LINE, 2019-2020) .
4.1 Notifications of removals or other enforcement decisions	There are no notifications of content removal.
4.2 Appeal processes against removals or other enforcement decisions	A user may appeal removal decisions through LINE's contact form.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Users can report any content that violates LINE's policies. Reports are reviewed by LINE's team and they 'take appropriate action' (LINE, n.d.) if they find any violations of such policies. In addition to responding to the user reports, LINE's monitoring system/team actively review the posted content by users (as described in Section 4 above). The marginal economic costs of using automated tools to identify TVEC are probably very low (although fixed costs may be substantial), whereas the marginal economic costs of using human moderators to this end are probably relatively high. LINE is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.

6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	LINE may delete content, or suspend or delete a user's account, without prior notice, if they believe that the user is violating or has violated its policies.
7. Does the service issue transparency reports (TRs) on TVEC?	No. However, LINE does issue TRs covering three matters: user information disclosure/deletion requests from law enforcement, actions taken against posts that violate LINE's ToS or applicable laws, and message and call encryption deployment status (LINE, 2019-2020).
8. What information/fields of data are included in the TRs?	In the report on the actions taken against violating posts on LINE services, LINE reports the number of content suspended, and percentages assigned to different categories, including Spam, obscene content, solicitation, unpermitted commercial use of accounts, disturbing and problematic content, promotion of illegal activity, and 'others'. TVEC seems to fall within the 'promotion of illegal activity' category (given the examples in Section 9 below), but this not explicitly stated.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	LINE clarifies that disturbing and problematic content may be 'excessively hateful remarks, photos of dead bodies, click fraud, links to phishing sites, etc.', and promotion of illegal activity may include 'announcements of attacks or bombings, sale of illegal drugs, selling online data (such as accounts, coins, and avatars) for real money, etc.'
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Unknown.

31. Huoshan

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There is no specific definition. However, Huoshan's ToS ban any content that promotes terrorism and extremism (not specifically violent extremism).
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.huoshanzhibo.com/agreement/

3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	No procedures are specified. Huoshan does inform that it keeps records of alleged violations of laws and regulations and suspected crimes, and report the same to the relevant competent authorities in accordance with the law, cooperating with any relevant investigations.
4.1 Notifications of removals or other enforcement decisions	No notifications are specified.
4.2 Appeal processes against removals or other enforcement decisions	No appeal processes are specified.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Users can report any type of unlawful activity or content on Huoshan. Huoshan's team of moderators reviews these reports and takes action accordingly. In addition, Huoshan has staff allocated to content moderation, and is increasing its efforts to improve its 'auditing standards' (Yoo, 2018). The marginal economic costs of using human moderators to detect objectionable content are probably relatively high. Huoshan is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	If a user violates Huoshan's ToS, Huoshan may delete posts or comments, restrict some or all of the functions of the infringer's account, or terminate access to its services.
7. Does the service issue transparency reports (TRs) on TVEC?	No.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.

10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Unknown.

32. Ask.fm

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No definition is provided. However, Ask.fm's Community Guidelines state that terrorist organisations and violent extremist groups that intend to encourage or commit terrorist or violent criminal activity are prohibited from maintaining a presence on Ask.fm to promote any of their campaigns or plans, celebrate their violent acts, fundraise, or recruit young people. The terms 'terrorist organisations' and 'violent extremist groups' are not defined. Additionally, users cannot post content that contains any threat of any kind, including threats of physical violence to themselves or others, or incites others to commit violent acts against themselves or others. No explicit definitions of the words "terrorist", "Terrorism" or "extremism" are provided.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://about.ask.fm/legal/2019-07/en/terms.html and https://about.ask.fm/community-guidelines/
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	Not applicable. Ask.fm does not offer any form of live stream capability.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	Ask.fm broadly states that they have the right to monitor users' access to or use of its services for violations of its ToS and to review or edit any content. Ask.fm also states that they can block or disable access to any content that they determine is objectionable or harmful to others, without prior notice.
4.1 Notifications of removals or other enforcement decisions	Content that violates Ask.fm's ToS or Community guidelines is removed, whereupon the owner receives a written warning. Ask.fm provides users with reasonable notice, if their access to the services and/or the profile is

	going to be suspended or terminated. The warning can be sent automatically by the system or manually by moderator several times before the actual profile ban.
4.2 Appeal processes against removals or other enforcement decisions	Users whose accounts have been banned may appeal this decision.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Users are able to report content that they believe violates Ask.fm's policies. Reports are sent to Ask.fm's team for review. Ask.fm asserts that they evaluate all reports. Ask.fm also states that they may access users' content and information when they believe it is reasonably necessary to enforce its ToS and protect the safety of Ask.fm's users or members of the public. The marginal economic costs of using human moderators to identify objectionable content are probably relatively high. Ask.fm is in onboarding to become a GIFCT member.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Violations of Ask.fm's ToS may lead to the suspension or termination of the infringer's account or access to Ask.fm's services.
7. Does the service issue transparency reports (TRs) on TVEC?	No.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Yes. It has been reported, for example, that one Ask.fm account offered advice on how to join ISIS fighters in Iraq, as well as what weapons one could expect to be equipped with on arrival. (Miller, 2014)

33. YY Live

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There is no specific definition. However, YY Live's ToS state that users cannot publish, transmit, disseminate, and store violent content, as well as content that promotes terrorism, extremism and related activities.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://zc.yy.com/license.html
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	No procedures are specified.
4.1 Notifications of removals or other enforcement decisions	No notifications are specified.
4.2 Appeal processes against removals or other enforcement decisions	No appeal processes are specified.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	No information is provided. However, research has shown that YY Live implements keyword censorship and surveillance. (Knockell, 2015) Specifically, to enforce its ToS, YY Live has a team within its data security department that maintains "24-hour surveillance" on content and is supported by a system that periodically "sweeps" the platform for offensive content and "automatically" filters keywords. (Knockell, 2015) The marginal economic costs of using automated tools to identify objectionable content are probably very low (although fixed costs may be substantial), whereas the marginal economic costs of using human moderators to this end are probably relatively high. YY Live is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.

6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	In case of violation of its ToS, YY Live may restrict or freeze the offender's use of their YY account, and restrict or suspend access to one or more specific products, services or functions (such as live video).
7. Does the service issue transparency reports (TRs) on TVEC?	No.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Unknown.

34. Twitch

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	Twitch updated its Terrorism and Extreme Violence guidelines in October, 2020 to be more explicit that it does not allow TVEC. Twitch does not allow content that depicts, glorifies, encourages, or supports terrorism, or violent extremist actors or acts. This includes threatening to or encouraging others to commit acts that would result in serious physical harm to groups of people or significant property destruction. This also includes displaying or linking to terrorist or extremist propaganda, including graphic pictures or footage of terrorist or extremist violence, even for the purposes of denouncing such content. (Twitch Community Guidelines, https://www.twitch.tv/p/en/legal/community-guidelines/) The clarifications of this update broadened the definition of content that fits in the Terrorism and Extreme Violence category (including forms of behaviour in this category that were previously categorised as other types of abuse), resulting in a substantial increase in enforcements (more on this in Section 9 below) of this kind (in percentage terms, if not in absolute number). Twitch's Community Guidelines also provide that acts and threats of violence will be taken seriously and are considered zero-tolerance violations. All accounts

	associated with such activities will be indefinitely suspended. This includes, but is not limited to: • Attempts or threats to physically harm or kill others • Use of weapons to physically threaten, intimidate, harm, or kill others. Twitch also prohibits hateful conduct, defined as any content or activity that promotes or encourages discrimination, denigration, harassment, or violence based on the following protected characteristics: race, ethnicity, color, caste, national origin, immigration status, religion, sex, gender, gender identity, sexual orientation, disability, serious medical condition, and veteran status. It also provides certain protections for age. Twitch has zero tolerance for hateful conduct, meaning it acts on every valid reported instance of hateful conduct. It affords every user equal protections under its policy, regardless of their particular characteristics.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.twitch.tv/p/en/legal/community-guidelines/, https://www.twitch.tv/p/en/legal/terms-of-service/, and https://help.twitch.tv/s/topic/0TO1U000000CjnZWAS/moderation-safety?language=en_US
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	Because Twitch is primarily a live streaming service, its terms and policies are designed for and are directly applicable to livestreamed content.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	Twitch takes enforcement action against accounts that violate its ToS and/or Community Guidelines. Twitch considers several factors when reviewing reports of violations, including signals of intent, surrounding context, potential harm to the community, legal obligations, and others. Depending on the nature of the violation, Twitch takes a range of actions that vary from issuing a warning, imposing a temporary suspension on the account, and for more serious or repeat offenses, an indefinite suspension. A warning is a courtesy notice. Twitch may also remove content associated with the violation. Repeating a violation for which a user has been already warned, or committing a similar violation, will result in a suspension. Temporary suspensions range from 24 hours to longer time periods that can exceed 30 days. If an account is suspended, the user may not access or use Twitch's services, including watching streams, broadcasting, chatting, creating other accounts and appearing/participating in a third party channel. After the

	suspension is complete, the user is able to use Twitch's services again. Twitch keeps a record of past violations, and multiple suspensions over time can lead to an indefinite suspension. For the most serious offenses, Twitch immediately and indefinitely suspends the account.
4.1 Notifications of removals or other enforcement decisions	Enforcement decisions are communicated via email to the uploader of the content. These include information on the type of content that was removed, a detailed explanation of where it happened and examples of the behaviors and pointing to the community guidelines (linking back to the applicable section of Twitch's Community Guidelines), as well as the duration of any penalties imposed as a result of the violation. These notifications also include information on how to appeal enforcement decisions. Note that Twitch will not notify the uploader of the content in cases involving illegal activity or where any notification may compromise any subsequent investigation by relevant authorities.
4.2 Appeal processes against removals or other enforcement decisions	If a user thinks that he or she did not violate Twitch's Community Guidelines, they may submit an appeal in response to an enforcement decision. In the appeal, the user must include the reason they believe the decision was incorrect. Once the appeal has been reviewed, Twitch notifies the user of the result.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Twitch makes available reporting tools at the service layer to enable users to report content or behaviour that violates Twitch's Community Guidelines, whether in the live broadcast, within the chat, or associated with a video file. At our service level Twitch also has proactive Machine Learning detection that flags content related to nudity, sexual content, gore and extreme violence which is flagged for review. Twitch blocks usernames that violate the Community Guidelines. Reports are reviewed by Twitch's Safety team, with reports of extreme violence and terrorist content receiving a priority. A second layer of moderation is made possible via Twitch's suite of tools that enables a channel owner (sometimes referred to as broadcaster) to designate other users as moderators of their channel. By doing so, those users then have the ability to ban bad users, block terms and phrases, require phone verification and remove messages from chat and take the same actions made available to the channel owner. Third, Twitch makes available to channel owners a tool that uses machine learning and natural language processing algorithms to prevent the display of messages within chat until they can be reviewed by a channel moderator before appearing to other viewers in the chat. This is referred to as "AutoMod" (Twitch, n.d.). AutoMod

	categories are focused on discrimination, sexual content, hostility and profanity. The marginal economic costs of using automated tools to identify objectionable content are probably very low (although fixed costs may be substantial), whereas the marginal economic costs of using human moderators to this end are probably relatively high. Twitch incurs no costs with regard to user moderators. Twitch is owned by Amazon, which joined the GIFCT in September 2019.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Violations of Twitch's Community Guidelines may lead to removal of content, a strike on the account, and/or suspension of the account. Serious offences are punished with immediate suspension.
7. Does the service issue transparency reports (TRs) on TVEC?	Yes. Twitch issued its first TR in February 2021, covering the whole of the year 2020 (Twitch, 2020). In the Section 'Reports and Enforcement' of its TR, TVEC-specific information is found. Twitch reports that it did not have any instances of live-streamed terrorist activity in 2020.
8. What information/fields of data are included in the TRs?	The following information is included for H1 and H2 of the reporting period: - The percentage of moderation coverage in channels by the tool AutoMod, channel moderators and both; - The number of manual and 'proactive' (i.e. with the aid of automated tools such as Blocked Terms and AutoMod) removals of chat messages by channel moderators; - The number of permanent channel bans and channel timeouts imposed by channel moderators; - The aggregate number of user reports for violations of the following categories: Terrorism, Terrorist Propaganda and Recruitment; Adult Nudity, Pornography and Sexual Conduct; Violence, Gore, Threats and Other Shocking Conduct; Hateful Conduct, Sexual Harassment and Harassment; and Viewbotting, Spam and Other Community violations; - The total number of enforcement actions; - The number of enforcement actions in the category Hateful Conduct, Sexual Harassment and Harassment; - The number of enforcement actions in the category Violence, Gore, Threats and other Shocking Content; - The number of enforcement actions in the category Adult Nudity, Pornography and Sexual

	Conduct; - The number of enforcement actions in the category Spam and other Community Guidelines Violations; and - The number of enforcement actions in the category Terrorism, Terrorist Propaganda and Recruitment, broken down into enforcements for showing terrorist propaganda and for glorifying or advocating acts of terrorism, extreme violence or large-scale property destruction.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Twitch explains that the vast majority of content removals on Twitch are removals of chat messages by channel moderators acting within individual channels. However, Twitch is a live-streaming service, and the vast majority of the content on Twitch is ephemeral. For this reason, Twitch does not focus on "content removal" as the primary means of enforcing streamer adherence to its Community Guidelines. Rather, live content is flagged by either machine detection or user reports, to Twitch's team of content moderation professionals, who then issue "enforcements" (typically a warning or timed channel suspension) for verified violations. If there happens to be recorded content that accompanies a violation, that content is removed. But most enforcements do not require content removal, because apart from the report, there is no longer a record of the violation - the live, violative content is already gone. For this reason, Twitch considers that the most appropriate measure of its safety efforts is 'enforcements' -hence the preponderance of this metric in its TR. For the sake of clarity, Twitch notes that the statistics regarding enforcements in the Section 'Reports and Enforcements' of its TR do not include, and are not duplicative of, the channel-level enforcements discussed in the Section 'Moderation in Channels: Coverage, Removals and Enforcements'.
10. Frequency/timing with which TRs are issued	On a yearly basis.
11. Has this service been used to post TVEC?	Yes. During a coordinated attack on Twitch's service in May 2019, certain users broadcasted offensive content, including past clips from the Christchurch attack. (Marshall, 2019) Also, In October 2019, a shooter in Halle, Germany livestreamed his attack on Twitch. (British Broadcasting Corporation (BBC), 2019) The attack was viewed by approximately 2,500 users before Twitch removed the footage of that attack, and it did not reappear on the service.

35. Tumblr

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There is no specific definition. However, Tumblr's Community Guidelines state that Tumblr does not tolerate content that promotes, encourages, or incites acts of terrorism. That includes content which supports or celebrates terrorist organisations, their leaders, or associated violent activities. The term 'terrorist organisations' is not defined. Also, Tumblr prohibits hate speech, understood as content that promotes or incites the hatred of, or dehumanizes, individuals or groups based on race, ethnic or national origin, religion, gender, gender identity, age, veteran status, sexual orientation, disability or disease.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.tumblr.com/policy/en/terms-of-service and https://www.tumblr.com/policy/en/community
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	If Tumblr concludes that a user is violating its policies, they may send the user a notice via email. If the user cannot explain or correct their behaviour, Tumblr may take action against their account. Tumblr notes that it reserves the right to suspend accounts, or remove content, without notice, for any reason, but particularly to protect its services, infrastructure, users, and community.
4.1 Notifications of removals or other enforcement decisions	There are no notifications of content removal.
4.2 Appeal processes against removals or other enforcement decisions	Users may contact Tumblr support to appeal a content removal decision.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Users can report any type of unlawful activity or content on Tumblr. Tumblr states that its trained experts review the reported content and take the 'appropriate action´. Reports do not always result in the content being removed. Sometimes Tumblr's experts determine that the reported content does not violate Tumblr's Community Guidelines.

	Tumblr does use automated tools to identify potentially TVEC-related content for human review, in addition to user reports. The marginal economic costs of using automated tools to detect objectionable content are probably relatively low (although fixed costs may be substantial), whereas the costs of using human moderators are likely relatively high. Tumblr is not a member of the GIFCT, but does participate in the GIFCT's Hash Sharing Consortium.[51]
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Tumblr may terminate or suspend the infringer's access to or ability to use any and all of Tumblr's services immediately, without prior notice or liability.
7. Does the service issue transparency reports (TRs) on TVEC?	No. Oath, previous controller of Tumblr (Alexander, 2019), does release transparency reports. Up until the year 2018, they included Tumblr. However, the reports are very broad and do not break down the information per company controlled by Oath (for example, government requests for removal of content included both Yahoo and Tumblr). Also, there is no information specific to TVEC (Verizon Media, 2019). In 2019 Tumblr was sold to Automattic. One Tumblr Transparency Report (covering the H2 2019) has been published ever since, but it concerns government requests for user data and content removal only. There is no information on TVEC (Tumblr, 2019)
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Yes. Tumblr is reportedly fraught with pages promoting Nazism, white supremacy, ethno-nationalism, and far-right terrorism (Barnes, 2019) (Fisher-Birch, 2018).

36. Flickr

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There is no specific definition. However, Flickr's ToS do prohibit posting content related to terrorism. Also, Flickr has a zero-tolerance policy towards attacking a person or group based on, but not limited to, race, ethnicity, national origin, religion, disability, disease, age, sexual orientation, gender, or gender identity.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.flickr.com/help/terms and https://www.flickr.com/help/guidelines
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	Whilst Flickr relies on a user moderation regime with regard to nudity and indecency, this system does not apply to TVEC, given that posting of TVEC leads to the deletion of the infringer's account. The criteria for identifying TVEC are not specified, though.
4.1 Notifications of removals or other enforcement decisions	No notifications are specified.
4.2 Appeal processes against removals or other enforcement decisions	No appeal processes are specified.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Users are able to report any content they consider violates Flickr's Community Guidelines. Flickr's staff review such reports to determine whether there is a violation, and take appropriate action. The marginal economic costs of using human moderators to detect objectionable content are probably relatively high. Flickr is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Posting TVEC content leads to the deletion of the relevant user's account. Flickr informs that they may report this conduct to law enforcement.

7. Does the service issue transparency reports (TRs) on TVEC?	No.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Yes. On Flickr, a virtual monument was created for foreign *jihadi* fighters killed in Syria, featuring their name, origin, and admiring remarks about their devoutness and combat strength (Weimann, 2014).

37. VK

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There is no specific definition. However, VK's ToS prohibit users from loading, storing, publishing, disseminating, making available or otherwise using any information that contains extremist materials and that promotes criminal activity or contains advice, instructions or guides for criminal activities. Similarly, VK's Platform Standards prohibits users from posting content which promotes illegal activities, criminal organisations or terrorism. Content that propagates and/or incites racial, religious, or ethnic hatred or hostility, including hatred or hostility towards a specific gender, orientation, or any other individual attributes or characteristics of a person (including those concerning a person's health) is also prohibited. VK follows the legal definition of terrorist or violent extremist content of the countries in which VK is present.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://vk.com/terms, https://vk.com/licence and https://m.vk.com/safety?lang=en&section=standarts

3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	All conditions from the VK ToS (https://vk.com/terms), License Agreement (vk.com/licence) and Community Standards (vk.com/safety?section=standards) apply to live-streamed content as well as all other types of content.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	No specific procedures are disclosed. VK broadly states that it reserves the right, at its own discretion as well as upon receipt of information from other users or third parties, to modify (moderate), block or remove any information published in breach of VK's ToS, or suspend, limit or terminate the infringer's access to all or any sections or services of VK at any time, with or without advance notice. Also, VK reserves the right to remove a user's personal page and/or suspend, limit or terminate the user's access to any of VK's services, if VK believes that the user poses a threat to VK and/or its users.
4.1 Notifications of removals or other enforcement decisions	Content removals are notified to users, even content listed in the Federal List of Extremist Materials of the Ministry of Justice of the Russian Federation.
4.2 Appeal processes against removals or other enforcement decisions	If a user disagrees with the decision to block or remove certain content, they can contact VK Support.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	VK uses a hybrid method of moderation. VK responds to reports from users, regulatory agencies and other organisations, also conducting internal monitoring through 'automatic search and inappropriate content removal mechanisms'. Examples of VK's automated tools is the use of digital fingerprints to quickly locate harmful content and neural networks. VK notes that the majority of 'dangerous content' is deleted before anyone even sees it (VK, 2020). Any person can report illegal, offensive, or misleading content with the help of the Report button. VK's moderation team reacts as quickly as possible to ban violators and block content that violates VK's rules or the applicable laws. Also, VK allows users to create 'Communities' and become administrators and moderators of them. According to VK's ToS, Community administrators and moderators bear liability for moderation and blocking of content uploaded to the pages that are under control of their communities. In particular, administrators and moderators must delete any content in breach of VK's ToS or applicable laws.

	The marginal economic costs of using automated tools to identify objectionable content are probably very low (although fixed costs may be substantial), whereas the marginal economic costs of using human moderators to this end are probably relatively high. VK is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Violations of VK's ToS including when creating and administering a Community entitle VK to remove/delete violating content, temporarily block the infringer's access to VK, exclude the content from search results or terminate the infringer's account.
7. Does the service issue transparency reports (TRs) on TVEC?	No. However, in VK's Safety Guidelines and Platform Standards VK reports a few metrics concerning the violation of its policies. There is no information on TVEC, though. See (VK, 2020) and (VK, 2020)
8. What information/fields of data are included in the TRs?	Number of pieces of content, profiles and communities blocked due to promotion of suicide, school violence (2019 statistics) hatred or hostility (Q1 and Q2 2020 statistics) and drug distribution (2019 statistics; https://vk.com/safety?section=health)
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not information provided.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Yes. ISIS accounts have been found in VK (Lokot, 2014).

38. Medium

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There is no definition. However, Medium's ToS provide that Medium does not allow content or actions that threaten, encourage, or incite violence against anyone, directly or indirectly; content that promotes violence or hatred against people based on characteristics like race, ethnicity, national origin, religion, disability, disease, age, sexual orientation, gender, or gender identity; posts or accounts that glorify, celebrate, downplay, or trivialize violence, suffering, abuse, or deaths of individuals or groups; and calls for intolerance, exclusion, or segregation based on protected characteristics. The glorification of groups that do any of the above is also prohibited.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://medium.com/policy/medium-rules-30e5502c4eb4 and https://medium.com/policy/medium-terms-of-service-9db0094a1e0f
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	For all user-reported content, Medium takes into account factors like newsworthiness, the context and nature of the posted information, reasonable likelihood, breadth, and intensity of foreseeable social harm, and applicable laws. In evaluating controversial and extreme content (not specifically violent extremist content) under Medium's Rules, moderators employed by Medium apply a risk analysis that includes, at a minimum, the following questions: - What are the foreseeable negative consequences of the information being propagated by Medium, and shared on other social media networks? - How severe might the potential impact be? - What is the likelihood of the negative consequence occurring? - Who will likely be affected as a result? - Is there information from nationally and internationally recognized institutions, (such as the CDC, WHO, and other official bodies) to help us determine if content presents an elevated risk? (Medium, n.d.) Medium provides the following examples of content areas with elevated risk, which is therefore more

	likely to be suspended or subject to reduced distribution: - Pseudo-scientific claims related to asserting the superiority or inferiority of a particular group (on bases including race, ethnicity or gender). - Conspiracy theories that have an associated history of harassment or violent incidents among adherents, or theories that may foreseeably incite or cause harassment, physical harm, or reputational harm. (Medium, n.d.)
4.1 Notifications of removals or other enforcement decisions	Upon investigating or disabling content associated with a user's account, Medium notifies the user, unless it believes the account is automated or operating in bad faith, or that notifying the user is likely to cause, maintain or exacerbate harm to someone.
4.2 Appeal processes against removals or other enforcement decisions	If a user believes his or her content or account has been restricted or disabled in error, or believes there is relevant context Medium was not aware of in reaching its determination, the user can file an appeal.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Users can flag content or accounts that violate Medium's Rules, or file a report containing a description of the alleged violation. Reported posts and users are reviewed by Medium's Trust & Safety team for Rules violations, after which appropriate actions are taken. The marginal economic costs of using human moderators to identify objectionable content are probably relatively high. Medium is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Violations of Medium's Rules may result in warnings, account restrictions, limited distribution of posts and content, suspension of content, and suspension of the violating account. Controversial and extreme content (again, not specifically violent extremist content) is particularly likely to be subject to suspended or limited distribution (Medium, n.d.).
7. Does the service issue transparency reports (TRs) on TVEC?	No. Medium issued a TR in 2015 (Medium, 2015) covering government requests for information or content removal in 2014, but there was no specific information on TVEC.

8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Unknown.

39. Odnoklassniki

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There is no specific definition. However, Odnoklassniki's ToS ban any propaganda or advocacy of hatred or supremacy based on social, racial, national or religious aspects; any content containing threats or inciting violence or criminal violations; and the publication of any information of extremist nature. The term 'extremist' is not defined.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://ok.ru/regulations
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	A few rules are specified at https://ok.ru/help/54/4532. Users can use OK Live anonymously, subject to functionality restrictions. To enjoy all functionalities, users must either use their Odnoklassniki profile or register a new profile using their phone number.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals and appeal processes against removal decisions?	Odnoklassniki broadly states that they may warn, notify or inform users of non-compliance with its ToS. The instructions provided by Odnoklassniki in these cases are mandatory for users. Also, Odnoklassniki explains that they may delete any content which in its opinion violates and/or may violate the applicable laws, its ToS, or cause harm or potential harm to, or threaten the safety of other users or third parties.
4.1 Notifications of removals	Odnoklassniki notifies users of their violations of its ToS at its discretion.

4.2 Appeal processes against removal decisions	No appeal processes are specified.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Users may become moderators of Personal Pages of other users, or create Groups and become administrators of them. In these cases, they have the obligation to moderate the content posted on such pages and groups. Users can also become moderators of videos and photos, by downloading the Odnoklassniki Moderator App (Odnoklassniki, n.d.). Users can report content that violates Odnoklassniki's ToS. Odnoklassniki's team reviews such reports and decides what actions to take. The marginal economic costs of using employed human moderators to detect objectionable content are probably relatively high. User moderators entail no cost for Odnoklassniki. Odnoklassniki informs that it does not perform and has no technical capability to perform automatic censorship of information in the publicly accessible sections of its Social Network or in the users' Personal Pages, or censorship of personal messages. Nor do they perform pre-moderation of information and content posted by users. Odnoklassniki is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Violation of Odnoklassniki's ToS give Odnoklassniki the right to suspend, restrict, or terminate the infringer user's access to its social network.
7. Does the service issue transparency reports (TRs) on TVEC	No.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Yes. TVEC content in support of IS has been found on Odnoklassniki (Clifford & Powell, 2019)

40. Haokan Video

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There is no definition. However, Haokan's ToS prohibit the use of its services to engage in illegal or improper activities, including the spreading of violence, murder and terrorism. The term 'terrorism' is not defined.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://haokan.baidu.com/videoui/page/about#agreement
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	Haokan broadly states that it reserves the right to block or remove content at any time without notice, in case Haokan determines there has been a violation of its policies.
4.1 Notifications of removals or other enforcement decisions	No notifications are specified.
4.2 Appeal processes against removals or other enforcement decisions	No appeal processes are specified.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Haokan provides no information in this regard. Haokan is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Haokan informs that violations of its ToS give Haokan the right to terminate or restrict the access to the infringer's account, and to delete any content violating its ToS, without prior notice.

7. Does the service issue transparency reports (TRs) on TVEC?	No.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Unknown.

41. Smule

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No definition is provided. However, Smule's Community Guidelines prohibit any content that promotes bigotry, discrimination, hatred, intolerance or racism; is hateful, offensive or shocking; or incites violence.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.smule.com/en/s/communityguidelines and https://www.smule.com/en/termsofservice
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	Smule broadly states that it does not pre-screen any user content, but reserves the right to remove or delete any content in its sole discretion, with or without notice, especially when the content violates its Community Guidelines or ToS. If Smule finds 'objectionable content', it takes appropriate action, including warning the user, suspending or terminating the user's account, removing all of the user's content, and/or reporting the

	user to law enforcement authorities, either directly or indirectly.
4.1 Notifications of removals or other enforcement decisions	There are notifications in the form of warnings, at Smule's discretion.
4.2 Appeal processes against removals or other enforcement decisions	No appeal processes are specified.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Users can report any content that violates Smule's ToS and Guidelines. Smule reviews the material flagged by Smule members and may remove it if is deemed inappropriate or unsafe for the Smule community, or if it otherwise violate Smule's Guidelines or ToS. The marginal economic costs of using human moderators to detect objectionable content are probably relatively high. Smule is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	If a user is found in violation of Smule's Guidelines or ToS, Smule may warn the user, remove any offending content, permanently terminate the user's account, notify law enforcement, or take legal action against the infringer.
7. Does the service issue transparency reports (TRs) on TVEC?	No.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Unknown.

42. KaKaoTalk

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	Kakao recently updated its operation policy that prohibits "posting any content that violates human dignity, incites violence, and instigates discrimination or prejudice due to reasons that include an individual's place of origin (including country and region), race, appearance, disability or illness, gender, gender identity, sexual orientation, and other factors associated with an individual's identity". In <Kakao's Commitment to End Online Hate Speech>, hate speech is defined as "offensive speech targeting a specific person or group of persons" on the basis of "actions of discrimination, incitement to prejudice, insult and social exclusion due to the reasons that include an individual's place of origin (including country and region), race, appearance, disability or illness, gender, gender identity, sexual orientation, and other factors associated with an individual's identity"
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.kakao.com/policy/oppolicy?lang=en (Article 3, paragraph 2, item 15)
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	Kakao TV applies wide and strict restriction on streaming illegal, violent or hateful content. Kakao TV's administrators are checking on all live-streaming contents in real time, and inform users that the service managers can immediately shut down the live streaming whenever the content violates the policy. Additionally, based on Kakao's Community Guidelines, the streaming contents in Kakao TV are subject to the discrimination banning rule, which bans all forms of discriminating expression or promotion on stereotypical perspective. Kakao TV is the only service of Kakao that provides users' livestreaming to public audiences.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	KakaoTalk broadly states that, in case of violation of its policies or applicable laws, it is able to investigate the breaches, delete the posts in question temporarily or permanently, or restrict all or part of its services temporarily or permanently. Whether the restriction is temporary or permanent depends on the accumulated number of violations; however, any explicit unlawful activities prohibited under applicable laws and regulations lead to permanent restriction, without delay, regardless of the accumulated number of violations.

4.1 Notifications of removals or other enforcement decisions	The enforcement actions above are notified to users via email or other means within the app, at the earliest convenience, except in case of urgent need to protect other users.
4.2 Appeal processes against removals or other enforcement decisions	Users can appeal the actions taken, and KakaoTalk informs appellants of the company's final decision after reviewing the appeal.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Users can create a 'story channel', become a master of it and invite managers to work in it. Masters and managers are administrators of story channels and act as moderators. Masters and managers can block and report users and content when they violate KaKaoTalk's policies. In addition, users can report any content that violates KaKaoTalk's policies. KaKaoTalk's team reviews these reports and takes appropriate action. Also, South Korean regulators, such as the National Policy Agency (NPA), the Communications Commissions, and the Korean Communications Standards Commission (KCSC) may request the deletion of any anti-social, violent and illegal information. Moreover, KaKaoTalk can apply restrictions for activities prohibited under its policies or in breach of applicable laws and regulations, without any report from users or regulators. Kakao monitors contents in story channels, including blogs and social media, based on keywords concerning TVEC and unlawful content. Kakao TV, Kakao's online video platform, is also subject to content monitoring, including live-streamed content. When problematic content is found on Kakao TV via monitoring, including TVEC, KaKao TV requires the uploader to alter (removing or revising the content) the content. If the content is not revised within 3 days, moderators delete the content and apply a temporary or lifetime ban in proportion to violent nature of the content and the user's aggregate number of violations. However, when it is decided that the content requires imminent action, moderators are authorised to instantly delete the post without delay. The marginal economic costs of using automated tools to identify objectionable content are probably very low (although fixed costs may be substantial), whereas the marginal economic costs of using human moderators to this end are probably relatively high. KaKaoTalk incurs no costs with regard to user moderators. KaKaoTalk is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.

6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	In case of violations of KaKaoTalk's policies, KaKaoTalk may issue a warning, delete the violating content, and temporarily or permanently restrict its services, depending on the accumulated number of violations. However, any explicit unlawful activities prohibited under the applicable laws and regulations lead to permanent restriction without delay, regardless of the accumulated number of violations.
7. Does the service issue transparency reports (TRs) on TVEC?	No. KaKaoTalk, however, does issue transparency reports (Daum Kakao, n.d.) disclosing the requests of South Korean government agencies to access user information, as well as content removals due to violation of its ToS and other policies, but there is no specific information on TVEC.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Twice a year.
11. Has this service been used to post TVEC?	Unknown.

43. DeviantArt

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No definition is provided. However, DeviantArt's ToS provide that commentaries that are overly aggressive or needlessly abusive are prohibited ('Prohibited Commentaries'). Moreover, users may not use DeviantArt for any unlawful purposes or to upload, post, or otherwise transmit any material that is unlawful, threatening, menacing, harmful or otherwise objectionable.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://about.deviantart.com/policy/service/, https://about.deviantart.com/policy/etiquette/ and https://about.deviantart.com/policy/submission/
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.

4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	After prohibited content is reported (a 'deviation'), the 'deviation owner' may receive an anonymous notification asking if the content is, for example, Mature Content, or whatever it was reported as. This gives the owner a chance to address and possibly remedy the situation. If the owner chooses not to take action and the content is not reported again, staff may agree that no deletion or tag is necessary, marking the report invalid. If the number of reports rises, however, it will rise in the staff's queue and they will more quickly take the appropriate action, whether that is adding a tag, deleting the content, or marking the report as invalid. It must be noted that even though a notification is sent to the deviation owner, every report still goes to DeviantArt's staff for final approval. This feature is simply a chance for a user to fix what might be an honest mistake (Kitsune, 2017). Use of any of the communication tools provided by DeviantArt for the purpose of deliberately aggressive or abusive behaviour can result in a disciplinary action (DeviantArt, n.d.). Forum threads that are misplaced, contain inappropriate subject matter, or contain an undesirable number of other violations of DeviantArt's policies are locked and closed to further commentary. As a registered member of DeviantArt, a user is able to participate as an administrator or member of a "Group", which is a set of user pages and applications formed for the purpose of collecting content, discussions and organising members of the site with common interests. Group administrators may determine its own rules and privileges for users who participate in the Group. As a general rule, DeviantArt will not interfere with Groups unless there is a clear violation of its policies. In these cases, DeviantArt can remove a Group and the Group's privileges. User accounts found to be demonstrating unacceptable behaviour, by failure to obey DeviantArt's policies or by engaging in abusive or disruptive community activity, can be subjected to a temporary account suspension (DeviantArt, n.d.). When an account is suspended, visitors to the suspended profile will be greeted by a "Suspended Account" message, which will be displayed instead of the normal profile page for the duration of the suspension. Administrative suspensions can be set for a variable period of time, with typical durations lasting for 24 hours, one (1) week, two (2) weeks, or thirty (30) days (one month). During this time, the profile will lose the ability to make posts, use most elements of the website, or interact with the community in general.

	The infringer receives notification of the action, which may include a private message or reason concerning why the action was taken, and a timer will be added to the relevant profile page. If the infringer is subject to further disciplinary action, previously recorded suspensions will be factored in. This may lead to a longer suspension or, in the case of repeat offenders, result in any new suspension being escalated to an account termination (DeviantArt, n.d.).
4.1 Notifications of removals or other enforcement decisions	If content is deleted by DeviantArt's staff, the owner gets a notification. Account suspensions are also notified.
4.2 Appeal processes against removals or other enforcement decisions	If the owner believes content is allowed on DeviantArt and the staff made a mistake, the owner can dispute the claim, explaining why. In this case, staff will give it a second consideration. Generally, DeviantArt allows its users to file appeals and make inquiries concerning content removals, violation notices, account suspensions and terminations or other administrative actions. Such appeals, inquiries and questions are reviewed and acted upon by DeviantArt's staff.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Group administrators are content moderators in their Groups. In addition, users can report any content that violates DeviantArt's policies. After a violation is brought to the attention of DeviantArt's staff, they review the report and take appropriate action. DeviantArt states that they have no ability to control the content users may upload, post or otherwise transmit using its service, and do not have any obligation to monitor such content for any purpose. The marginal economic costs of using employed human moderators to detect objectionable content are probably relatively high. User moderators entail no cost for DeviantArt. DeviantArt is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Violations of DeviantArt's policies may lead to a warning, deletion of content, account suspension or termination of the violator's membership, at DeviantArt's sole discretion.

7. Does the service issue transparency reports (TRs) on TVEC?	No.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Yes, Neo-Nazi groups have used DeviantArt to upload propaganda and recruit new members (Hayden, Mysterious Neo-Nazi Advocated Terrorism for Six Years Before Disappearance, 2019).

44. Meetup

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No definition is provided. However, according to Meetup's ToS, gratuitously graphic or violent content is prohibited; behaviour that incites violence against individuals or groups of people based on who they are or their beliefs is prohibited; and using Meetup to promote, facilitate, or organise violent, criminal, or non-consensual actions that endanger anyone, physically, mentally or emotionally, is also prohibited. Moreover, 'Groups' (sections within Meetup focused on specific interests or activities) must not contain content or promote events that organise, promote, provide for, distribute services for, or recruit for terrorist organisations; contain content or promote events that could threaten public or personal safety, including advocating for, inciting, or making aspirational statements or threats to commit violence against any group of people, individual person, or specific location, weapons and explosive-making, and calls for violence in response to private or public events.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://help.meetup.com/hc/en-us/articles/360002897532-Usage-and-content-policies-Rules-for-using-Meetup, https://help.meetup.com/hc/en-us/articles/360004285732-Meetup-social-media-community-standards,

	https://help.meetup.com/hc/en-us/articles/360002897712-Meetup-groups-and-events-policies and https://help.meetup.com/hc/en-us/articles/360027447252-Terms-of-Service
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	Meetup broadly states that violations of its policies and ToS may lead to the modification, suspension or termination of the infringer's account or access to Meetup, and when this happens, Meetup notifies the infringer of the reasons for the modification, suspension, or termination.
4.1 Notifications of removals or other enforcement decisions	Enforcement decisions are notified to users.
4.2 Appeal processes against removals or other enforcement decisions	If a user believes the modification, suspension, or termination has occurred in error, he or she can appeal the decision.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Group administrators are content moderators in their Groups, and have the ability to modify, suspend, or terminate users' access to the Groups they moderate. In addition, users can report any content that violates Meetup's policies. Meetup's Trust and Safety team reviews all reports and takes appropriate action. The marginal economic costs of using employed human moderators to detect objectionable content are probably relatively high. User moderators entail no cost for Meetup. Meetup states that they generally do not review content before it is posted (Meetup, 2019). Meetup is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Infringement of Meetup's policies may lead to content deletion, modification, suspension or termination of the infringer's account.

7. Does the service issue transparency reports (TRs) on TVEC?	No. Meetup does issue transparency reports (Meetup, 2017) that disclose government requests for access to users' information and requests for content removal based on Intellectual Property rights infringements, but there is no information on TVEC.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Unknown.

45. 4chan

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There is no definition. However, 4chan's ToS prohibit content that violates local or United States laws.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at http://www.4chan.org/rules#global4
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	According to 4chan, threads expire and are pruned by 4chan's software at a relatively fast rate. Since most boards are limited to ten pages, content is usually available for only a few hours or days before it is removed. Usually, missing posts are probably pruned automatically; however, in some cases they may have been removed by a moderator or 'janitor'. Moderators are individuals selected to perform general site maintenance. They may delete posts globally, ban users, close threads and carry out associated actions.

	Janitors are a class between 'end user' and 'moderator'. They are given access to 4chan's report system and may delete posts on their assigned board(s), as well as submit ban requests. Janitors are selected via an application, orientation, and testing process. Admission to the moderation team is by invitation only. The janitor program is occasionally opened to new applicants. There is no public record of content deletion and because threads are frequently pruned, there is no way of knowing which pieces of content have been removed by the moderation team. In short, there is no way for an end user to judge accurately the amount of moderation taking place at any given point in time. The 4chan moderation team reserves the right to block or ban access and remove content for any reason without notice. Users are temporarily blocked from posting when there is a pending ban request placed on their IP address. This block lasts 15 minutes from the time a janitor submits a ban request and is removed immediately if the request is denied by a moderator. If the request is approved, a regular ban is applied.
4.1 Notifications of removals or other enforcement decisions	No notifications are specified.
4.2 Appeal processes against removals or other enforcement decisions	Users can appeal bans it they believe an error has been made, by contacting the moderators.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	4chan states that it encourages reporting posts for review (4chan, n.d.). Moderators review the reported content and take appropriate action. The marginal economic costs of using employed human moderators to detect objectionable content are probably relatively high. User moderators entail no cost for 4chan. 4chan is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Breaking 4chan's Rules may result in post deletion, a temporary ban, or in some cases, permanent banishment.

7. Does the service issue transparency reports (TRs) on TVEC?	No.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Yes. For example, Neo-Nazi propaganda is common on 4chan (Arthur, 2019).

46. Google Drive

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There is no specific definition of TVEC. However, Google's Abuse Program Policies (Google, n.d.), which apply to Google Drive, have specific provisions on Violence, Hate Speech and Terrorist Activities. *Violence*: Users may not threaten to cause serious physical injury or death to a person, or rally support to physically harm others. In cases where there is a serious and imminent physical threat of injury or death, Google may take action on the content. Posting violent or gory content that is primarily intended to be shocking, sensational, or gratuitous is prohibited. If posting graphic content in a news, documentary, scientific, or artistic context, users must provide enough information to help people understand what is going on. In some cases, content may be so violent or shocking that no amount of context will allow that content to remain on Google's platforms. Also, users may not encourage others to commit specific acts of violence. *Hate speech*: Hate speech is not allowed. Hate speech is content that promotes or condones violence against or has the primary purpose of inciting hatred against an individual or group on the basis of their race or ethnic origin, religion, disability, age, nationality, veteran status, sexual orientation, gender, gender identity, or any other characteristic that is associated with systemic discrimination or marginalization. *Terrorist activities*: Google does not permit terrorist organizations to use Drive for any purpose, including recruitment. Google also strictly prohibits content related to terrorism, such as content that promotes terrorist acts, incites

	violence, or celebrates terrorist attacks. The term 'terrorist organizations' is not defined. If users post content related to terrorism for an educational, documentary, scientific, or artistic purpose, they must provide enough information so viewers understand the context.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.google.com/drive/terms-of-service/ and https://support.google.com/docs/answer/148505?visit_id=637064013896463652-1393240150&rd=1
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	When files are flagged for a violation, the owner of the file may see a flag next to the filename and he or she will not be able to share it. The file will no longer be publicly accessible, even to people who have the link. Users can request that their file be reviewed if they do not think it violates Google's ToS or program policies (Google, n.d.). If a user materially or repeatedly violates Google Drive's ToS or Program Policies, Google may suspend or permanently disable that user's access to Google Drive. Google gives prior notice in such cases. However, Google may suspend or disable a user's access to Google Drive without notice if he or she is using Google Drive in a manner that could cause Google legal liability or disrupt other users' ability to access and use Google Drive.
4.1 Notifications of removals or other enforcement decisions	No notifications are specified.
4.2 Appeal processes against removals or other enforcement decisions	No appeal processes are specified.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Users can report content that violates Google Drive's ToS and policies. Reports are assessed by Google's staff. Google states that reports do not guarantee removal of the file or any other action on Google's part. This is because content that a user disagrees with or deems inappropriate is not always a violation of Google's ToS or program policies. Google also indicates that they may review users' conduct and content in Google Drive for compliance with the ToS and Program Policies (Google, 2019). Google has reported that files in Google Drive are policed by an algorithm that looks out for abuse of its policies and automatically blocks files that are deemed to violate them. This system involves no human review (Titcomb, 2017). The marginal economic costs of using automated tools to identify objectionable content are probably very low (although

	fixed costs may be substantial), whereas the marginal economic costs of using human moderators to this end are probably relatively high. GoogleDrive is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Abusive material in violation of Google's ToS or other policies entitles Google to: - Remove the file from the account - Restrict sharing of a file - Limit who can view the file - Disable access to one or more Google products - Delete the Google Account (Google, n.d.)
7. Does the service issue transparency reports (TRs) on TVEC?	No. Google issues TRs (Google, n.d.) encompassing Google's products and services, including Google Drive. These reports contain a section on government requests to remove content based on violations of local laws or Google's ToS or policies, but there is no TVEC-specific information.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Yes. ISIS content has been found on Google Drive (Katz, To Curb Terrorist Propaganda Online, Look to YouTube. No, Really., 2018).

47. Dropbox

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No definition is provided. However, Dropbox's Acceptable Use Policy provides that users cannot use Dropbox to publish or share materials that contain extreme acts of violence or terrorist activity, including terrorist propaganda. Using Dropbox to advocate bigotry or hatred against any person or group of people based on their race, religion, ethnicity, sex, gender identity, sexual orientation, disability or impairment is also prohibited.

2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.dropbox.com/terms and https://www.dropbox.com/terms#acceptable_use
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	Not applicable.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	Dropbox states that if a user breaches the ToS or uses Dropbox's services in a manner that would cause a real risk of harm or loss to Dropbox or other users, Dropbox has the right to suspend or terminate the user's access. If Dropbox provides the user with advance notice, Dropbox will provide the user with an opportunity to export his or her content. If after such notice the user fails to take the steps Dropbox requires, Dropbox will terminate or suspend the user's access to Dropbox's services. Dropbox does not provide advance notice when a user is in material breach of the ToS, when doing so would cause Dropbox legal liability or compromise its ability to provide its services to other users, or when Dropbox is prohibited from doing so by law.
4.1 Notifications of removals or other enforcement decisions	No notifications are specified.
4.2 Appeal processes against removals or other enforcement decisions	Users can request that Dropbox review their takedown decision if they believe the content doesn't violate Dropbox's ToS.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Users and others, including trusted flaggers and non-governmental organizations, can report content that violates Dropbox's ToS and policies. Dropbox's team reviews these reports, investigates the alleged violation, and takes appropriate action. Dropbox also uses automated detection technology and employs a team of human reviewers. Dropbox has reported that its staff, on rare occasions, need to access users' file content, particularly to enforce its ToS and policies (Dropbox, n.d.). The marginal economic costs of using human moderators to identify objectionable content are probably relatively high. Dropbox is a member of the GIFCT.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Violation of Dropbox's ToS or other policies may lead to the loss of services on Dropbox, suspension or termination of the infringer's account.

7. Does the service issue transparency reports (TRs) on TVEC?	No. Dropbox issues TRs (Dropbox, n.d.) that contain a section on government requests to remove content based on violations of local laws or Dropbox's ToS or policies, but there is no TVEC-specific information.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Yes. ISIS content has been found on Dropbox (Bennett, 2019).

48. Microsoft OneDrive

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No definition is provided. However, Microsoft's Services Agreement (SA), which governs OneDrive, prohibits any activity that is harmful to others, such as posting terrorist or violent extremist content, communicating hate speech or advocating violence against others. Microsoft has stated that for the purposes of its services, they consider terrorist content to be material posted by or in support of organizations included on the Consolidated United Nations Security Council Sanctions List (United Nations Security Council) that depicts graphic violence, encourages violent action, endorses a terrorist organization or its acts, or encourages people to join such groups. The U.N. Sanctions List includes a list of groups that the U.N. Security Council considers to be terrorist organizations (Microsoft, 2016). In its Digital Safety Content Report (Microsoft, 2021), Microsoft clarifies that 'both terrorist and violent extremist content is prohibited on Microsoft platforms and services', and that Microsoft Services Agreement Code of Conduct prohibits the 'posting of terrorist or violent extremist content.'
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.microsoft.com/en-us/servicesagreement/
3. Are there specific provisions applicable to livestreamed content in	Not applicable.

the ToS or Community Guidelines/Standards?	
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	Microsoft states that it reserves the right to remove or block a user's content from OneDrive at any time if it is brought to its attention that the content may violate applicable law or its SA. When investigating alleged violations of its SA, Microsoft reserves the right to review the user's content in order to resolve the issue. However, Microsoft clarifies that it does not monitor OneDrive. Microsoft follows a "notice-and-takedown" process for removal of prohibited content, including terrorist content, which is to say that the "notice" is sent to Microsoft (by a government or a user, for example) and then Microsoft takes down the content. Thus, when the presence of terrorist content on Microsoft's hosted consumer services, including OneDrive, is brought to the company's attention via Microsoft's online reporting tool, Microsoft will remove it (Microsoft, 2016). As described in Microsoft's Services Agreement, "If you violate these Terms, we may stop providing Services to you or we may close your Microsoft account. We may also block delivery of a communication (like email, file sharing or instant message) to or from the Services in an effort to enforce these Terms or we may remove or refuse to publish Your Content for any reason. When investigating alleged violations of these Terms, Microsoft reserves the right to review Your Content in order to resolve the issue."
4.1 Notifications of removals or other enforcement decisions	Notifications are at Microsoft's discretion. Microsoft's Services Agreement states: "When there's something we need to tell you about a Service you use, we'll send you Service notifications. If you gave us your email address or phone number in connection with your Microsoft account, then we may send Service notifications to you via email or via SMS (text message), including to verify your identity before registering your mobile phone number and verifying your purchases. We may also send you Service notifications by other means (for example by in-product messages)."
4.2 Appeal processes against removals or other enforcement decisions	Microsoft's Account suspension appeals form is available at: https://www.microsoft.com/en-us/concern/AccountReinstatement
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Microsoft states that the Microsoft Services Agreement Code of Conduct prohibits the "posting [of] terrorist or violent extremist content." Microsoft encourages the reporting of content posted by – or in support of – a terrorist organization that depicts graphic violence, encourages violent action, endorses a terrorist organization or its acts, or encourages people to join such groups. Microsoft reviews these reports; takes action on content; and, if necessary, suspends accounts associated with violations of our Code of Conduct.

	In addition, Microsoft leverages a variety of tools, including hash-matching technology and other forms of proactive detection, to detect terrorist and violent extremist content. Microsoft uses scanning technologies (e.g., PhotoDNA or MD5) and other AI-based technologies, such as text-based classifiers, image classifiers, and the grooming detection technique to detect TVEC (Microsoft, 2021) The marginal economic costs of using human moderators to identify objectionable content are probably relatively high. Microsoft is a founding member of the GIFCT and participates in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	If a user posts content that is prohibited or otherwise materially violates the SA, Microsoft may take action against the user, including stopping access to OneDrive, closing the user's Microsoft account immediately, or blocking delivery of a communication (like email, file sharing or instant messaging) to or from the OneDrive. Microsoft may also block or remove infringing content. See also Section 4 above, and this 2016 blog entry: "**Observing notice-and-takedown**: We will continue our 'notice-and-takedown' process for removal of prohibited, including terrorist, content. When terrorist content on our hosted consumer services is brought to our attention via our online reporting tool, we will remove it. All reporting of terrorist content – from governments, concerned citizens or other groups – on any Microsoft service should be reported to us via this form." (https://blogs.microsoft.com/on-the-issues/2016/05/20/microsofts-approach-terrorist-content-online/)
7. Does the service issue transparency reports (TRs) on TVEC?	Yes. TVEC numbers for Skype are included in Microsoft's Digital Safety Content Report (Microsoft, 2021). This report is inclusive of Microsoft consumer products and services including (but not limited to) OneDrive, Outlook, Skype, Bing and Xbox. It must be noted that TVEC metrics are reported on aggregate for all Microsoft consumer services and products, and not on a per-product basis.
8. What information/fields of data are included in the TRs?	• Pieces of TVEC actioned • Number of accounts suspended due to TVEC • % of TVEC actioned that Microsoft detected • % of accounts suspended for TVEC that were reinstated upon appeal
9. Methodologies for determining/calculating/estimating	"Content actioned" refers to when Microsoft removes a piece of user-generated content from its products and services and/or blocks user access to a piece of user-generated content.

the information/data included in the TRs	"Account suspension" means removing the user's ability to access the service account either permanently or temporarily "Proactive detection" refers to Microsoft-initiated flagging of content on its products or services, whether through automated or manual review.
10. Frequency/timing with which TRs are issued	Not reported
11. Has this service been used to post TVEC?	Yes. ISIS videos have been hosted on OneDrive (Counter Extremism Project, 2018).

49. WordPress.com

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No definition is provided, though WordPress.com's ToS provide that WordPress.com does not allow websites of terrorist groups recognised by the United States government. The U.S. Department of the Treasury's Office of Foreign Assets Control maintains a list of "Specially Designated Nationals" (US Treasury, 2020), with which WordPress.com is prohibited by law from doing business. WordPress.com does not allow individuals, groups, or entities on that list to use WordPress.com (Word Press, n.d.). Genuine calls to violence are also prohibited. This include the posting of content which threatens, incites, or promotes violence, physical harm, or death, threats targeting individuals or groups, as well as other indiscriminate acts of violence.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://en-gb.wordpress.com/tos/ and https://en.support.wordpress.com/user-guidelines/
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	Not applicable.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or	WordPress.com has worked in conjunction with experts on online extremism, as well as law enforcement, to develop policies to address extremist (not specifically violent extremist) and terrorist propaganda. WordPress.com suspends websites that call for violence or that are connected to officially banned terrorist groups (per the US Treasury's OFAC list), regardless of content. WordPress.com also

other enforcement decisions and appeal processes against them?	implements other measures short of removal—for example, it may flag content and remove a site from the WordPress.com Reader, making the site's content more difficult to find. Flagging a site also removes it from all advertising programs run by WordPress.com. According to WordPress.com, one important way that extremist (again, not specifically violent extremist) sites are brought to its attention is through reports from dedicated government Internet Referral Units (IRUs). These organisations have expertise in online propaganda that private technology companies are not able to develop on their own. They work to identify sites that are being used by known terrorists to spread propaganda or to organise acts of violence. They report terrorist sites to WordPress.com using a dedicated email address that allows WordPress.com to more easily identify reports coming from a trusted source. WordPress.com does not automatically remove websites from WordPress.com. Rather, a human member of its Risk & Safety team reviews each report and makes a decision on whether it violates its policies. One important reason it reviews each report is to guard against the removal of material posted to legitimate sites (news organisations, academic sites) that discuss terrorism or a terrorist group. WordPress.com hosts sites for a number of very large news organisations, news bloggers, academics, and researchers who all publish legitimate reporting on terrorism. In another context, though, some of the materials they publish may qualify as terrorist propaganda, and if so, would be removed under WordPress.com' policies. WordPress.com states that context is very important and they cannot outsource these important decisions affecting legitimate online speech to a robot. Also, since the volume of reports it receives is not high relative to other online platforms, it is able to use more human, versus automated review, when acting on reports (Clicky, 2017).
4.1 Notifications of removals or other enforcement decisions	WordPress.com states that, depending on the scenario, it will email or add a warning notification in the dashboard of a user violating its policies. The notification will contain a link that the user can use to contact WordPress.com regarding the issue. However, those 'scenarios' are not specified (WordPress.com, n.d.).
4.2 Appeal processes against removals or other enforcement decisions	Users can appeal WordPress.com's enforcement actions when the users believe that the actions were taken in error. A real person will review the request and reply with a decision as soon as possible.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	WordPress.com does not pre-screen the content users post. Users are able to report content or sites in violation of WordPress.com's policies. In addition, as noted above, IRUs report terrorist and extremist sites to WordPress.com.

	WordPress.com evaluates those reports and takes appropriate action. The marginal economic costs of using human moderators to identify objectionable content are probably relatively high. WordPress.com is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	If WordPress.com finds a site or any of a site's content to be in violation of its policies, WordPress.com will remove the content, disable certain features on the account, and/or suspend the site entirely.
7. Does the service issue transparency reports (TRs) on TVEC?	Yes. Automattic (WordPress.com' parent company) issues TRs that contain a section on reports from IRUs relating to extremist (not specifically violent extremist) content (Automattic, n.d.). The last TR included data from 1 January to 30 June 2020.
8. What information/fields of data are included in the TRs?	- Number of IRU extremist (not specifically violent extremist) content notices - Number of notices for which sites/content were removed as a result - Percentage of notices for which sites/content were removed as a result The figures are broken down by month (January to June and July to December) and by reporting entity or country. Also, in the Summary section of its TR, Automattic reports the number of sites/content specified in the IRU notices for the period between 1 January 2018 – 30 June 2020.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	No information available.
10. Frequency/timing with which TRs are issued	On a half-yearly basis. Automattic has issued TRs for the following periods: - 2017: 1 Jul – 31 Dec - 2018: 1 Jan – 30 Jun - 2018: 1 Jul – 31 Dec - 2019: 1 Jan – 30 Jun - 2019: 1 Jul – 31 Dec - 2020: 1 Jan – 30 Jun
11. Has this service been used to post TVEC?	Yes. See Section 7 above.

50. Wikipedia

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No definition is provided. However, the Wikimedia Foundation's ToS, which govern Wikipedia, prohibit harassment, threats, stalking, and vandalism, among other things. The ToS also prohibit using Wikimedia's services in a manner that is inconsistent with applicable law.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://foundation.wikimedia.org/wiki/Terms_of_Use/en and https://en.wikipedia.org/wiki/Wikipedia:Policies_and_guidelines#Enforcement
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	Not applicable.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	The Wikipedia community has the primary role in creating and enforcing its policies. The community is composed of: - *Editors*: volunteers who write and edit the pages of Wikipedia - *Stewards*: volunteer editors tasked with the technical implementation of community consensus, with Checkuser (Wikipedia, 2019) and oversight (Wikipedia, 2020) powers. - *Bureaucrats*: volunteer editors with the technical ability (user rights) to promote other users to administrator or bureaucrat status, remove the admin status of other users, and grant and revoke an account's bot status. - *Administrators*: editors who have been trusted with access to restricted technical features ("tools"). For example, administrators can protect and delete pages, and block other editors (Wikipedia, 2020). Wikipedia's core content policies are: 1. Neutral point of view: All Wikipedia articles and other encyclopaedic content must be written from a neutral point of view, representing significant views fairly, proportionately and without bias. 2. Verifiability: It means that people reading and editing the encyclopaedia can check that information comes from a reliable source. 3. No original research: Wikipedia does not publish original thought. All material in Wikipedia must be attributable to a reliable, published source (Wikipedia, 2019).

	Content is deleted by the administrators if it is judged to violate Wikipedia's content or other policies, or the laws of the United States (Wikipedia, 2020). The deletion process encompasses the processes involved in implementing and recording the community's decisions to delete pages and media (Wikipedia, 2020). Normally, a deletion discussion must be held to form a consensus to delete a page. In general, administrators are responsible for closing these discussions, though non-administrators in good standing may close them under specific conditions. However, editors may propose the deletion of a page if they believe that it would be an uncontroversial candidate for deletion. In some circumstances, a page may be speedily deleted if it meets strict criteria set by consensus, which include pages that disparage, threaten, intimidate or harass their subject or some other entity, and serve no other purpose (Wikipedia, 2020). The Wikimedia Foundation states that it rarely intervenes in community decisions about policy and its enforcement. However, when the community requires intervention, or to address an especially problematic user because of significant disturbance or dangerous behaviour, the Wikimedia Foundation may investigate the user's use of the service (a) to determine whether a violation of any policies or laws has occurred, or (b) to comply with any applicable law, legal process, or appropriate governmental request. After the investigation, sanctions may be applied (see Section 6 below).
4.1 Notifications of removals or other enforcement decisions	Not applicable.
4.2 Appeal processes against removals or other enforcement decisions	Not applicable.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Editorial control, and therefore the detection of content that violates Wikipedia's policies, is in the hands of the Wikipedia community. Also, readers (Wikipedia users who do not make contributions) can contact Wikipedia's Volunteer Response Team to report any issue with content on available on Wikipedia. The Wikimedia Foundation states that it does not take an editorial role with respect to its projects, including Wikipedia. This means that it 'generally' does not monitor or edit the content of its projects' websites (Wikimedia Foundation, 2019). The Wikimedia Foundation incurs no costs with regard to Wikipedia community moderators. Wikipedia is not a member of the GIFCT, and does not participate in GIFCT's Hash Sharing Consortium.

6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	The Wikipedia community may issue a warning, investigate, delete pages created by, block, and/or ban users who violate the community's policies. The Wikimedia Foundation may refuse, disable, or restrict access to the contribution of any user who violates its ToS, ban a user from editing or contributing or block a user's account or access for actions violating its ToS, and take legal action against users who violate its ToS (including reports to law enforcement authorities).
7. Does the service issue transparency reports (TRs) on TVEC?	No. The Wikimedia Foundation does issue TRs (Wikimedia Foundation, n.d.) covering requests for user data and requests for content alteration and takedown, but there is no section specifically addressing TVEC.
8. What information/fields of data are included in the TRs?	In the section 'Requests for user information', under the heading 'emergency disclosures', the Wikimedia Foundation discloses the number of disclosures of user data in connection with terrorist threats. The Wikimedia Foundation proactively contacts law enforcement authorities when it becomes aware of troubling statements on Wikimedia projects, such as bomb threats. This does not amount, however, to removals of TVEC.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Unknown.

Annex C - Glossary

The following definitions and explanations are meant to clarify certain terms that are common in transparency reporting on TVEC.

Actioning accounts – In addressing TVEC-related issues, a content-sharing service may take action in response to the TVEC-related online activity of a user or an account. This could be endorsing or rewarding positive user behaviour, such as helpfully flagging or reporting problematic content. Conversely, it could be action to prevent or address negative user behaviour, such as sharing TVEC that violates the guidelines. Examples of the latter type of action include:

Banning – Banning a user prohibits them from logging on to a content-sharing service and/or from creating and using any new accounts.

Disabling/de-activating/suspending – Disabling an account — which could include removing, deleting, de-activating or suspending an account — is effectively closing an account which has violated guidelines. This may be temporary or permanent and may be open to redress mechanisms or subject to a specific period of time. It may or may not affect the accessibility of the account's past contributions on the content-sharing service, and may or may not be subject to an obligation to preserve data for law enforcement or similar purposes.

Reporting to law enforcement – A user or account may be reported to a law enforcement agency in order to address illegal activity or imminent risks to safety.

Restricting user privileges – An account may remain operable but with specific privileges restricted, muted, suspended or removed. These privileges may include the ability to live-stream, comment or post.

Warning – A warning message or notice may be issued to an account that has violated company guidelines.

Actioning content – Once the appropriate moderation outcome is determined, the content either remains on the online platform in its original state as is, or is actioned in some way by the moderator (company staff, technology and/or a designated third party). Action may also be taken on an interim basis while a moderation outcome is pending. Content may be actioned in a number of ways. These include:

Blocking/disabling – Blocking/disabling means restricting or removing access to specific content for a particular user or group of users. Geo-blocking, for example, restricts access to content for users whose IP addresses are registered within a specific physical location. The content may remain available to some users under specific circumstances.

De-listing – De-listing is the removal by a content-sharing service or user of content from recommendation lists for users, or from indexing within the 'explore' or 'discover' functions that allow users to search content on the content-sharing service.

De-monetising – De-monetising content is restricting its ability to leverage the content-sharing service's monetisation features. For example, de-monetising could involve removing the possibility for advertisements to appear alongside content that does not comply with relevant guidelines (e.g. content or others).

Down-ranking – Down-ranking allows content to remain available on the content-sharing service but with reduced visibility. Down-ranking is also known as down-listing, de-prioritising or limiting visibility.

Hiding/quarantining -- Notifications provided before content can be accessed are also known as interstitial notices. Content hidden behind an interstitial notice may become accessible to a user if specific conditions are met — such as users declaring their age or acknowledging that content may be offensive. Content may also be quarantined or hidden behind a notification to indicate that it is not accessible to users because it is under review or is in violation of a company's guidelines.

Notification – A moderator may add a notification to user-generated content, to make other users aware that it may be sensitive, disturbing, false, inappropriate for younger users, or otherwise challenging to community expectations, even though it may not violate company guidelines.

Removing – Removing is the process of a content-sharing service taking down content so it is no longer accessible to any users. The permanency of removal is determined by the content-sharing service's guidelines and redress mechanisms, and the legality of the content.

Appeals and reviews – A process by which one or more users who believe the outcome of a moderation decision is incorrect may seek reconsideration of that decision. Some content-sharing services that provide options for appeal or review may use automated review and/or human review. The review may be conducted internally by the service and/or by appropriate circumstances that involve members of the user community, or by an external, independent body, including the judicial authorities in respective countries. If a review results in a decision to reverse, overrule or change the initial moderation outcome, common forms of redress or resolution include restoring content or an account, actioning content (see above) or actioning an account (see above).

Banning – See Actioning accounts.

Blocking – See Actioning content.

Company guidelines – Company guidelines are also known as community standards, rules, acceptable use policy, terms of service or terms of use. These guidelines are commonly understood to be a set of expectations for what content or activity is or is not allowed on a company's service or product. These guidelines may also outline the actioning of content or accounts and user notification and redress mechanisms.

Content-sharing services – Content-sharing services are any online services that enable the transfer and dissemination of content, in whatever form, whether one-to-one, one-to-few or one-to-many.

De-activating – See Actioning accounts.

De-listing – See Actioning content.

De-monetising – See Actioning content.

Detection and moderation – Detection and moderation can occur at different stages and can take a number of forms. They may occur nearly simultaneously (for example, through automated systems) or sequentially over a period of time (for example, through human review of content reported by a user). The following reflect some common forms and definitions of detection and moderation.

Detection – Detection is the process of identifying TVEC or TVEC-related online activity on a content-sharing service. Detection may be:

1. Proactive – Proactive detection occurs when TVEC or TVEC-related online activity is detected as a result of company-led routine detection. Proactive detection can happen from human, tooling or hybrid systems of review established by a content-sharing service. Proactive detection can be:
    a. Proactive at upload – Proactive detection at upload occurs as soon as a user attempts to add TVEC to, or take specific TVEC-related online actions on a content-sharing service and **before** it is shared with or becomes accessible to others. This is primarily done by automated tools. Once such content or activity is flagged, various moderation actions can take place. For example, if the content is not obviously or overtly against guidelines, it might trigger a triage to human review.
    b. Proactive after upload – Proactive detection after upload occurs after TVEC has been added to a content-sharing service. Depending on the circumstances, this detection may occur **before or after** TVEC has been shared with or become accessible to other users. Again, once TVEC is flagged, various moderation actions can take place.

2. Reactive – Reactive detection occurs when TVEC or TVEC-related online activity is identified through a third-party report made to the content-sharing service. TVEC or TVEC-related online activity may be reported by users (see online community reports below) or by others, such as civil society organisations, governments, law enforcement, trusted notifiers, regulatory bodies, industry bodies, etc. Reports from government institutions or public authorities may take the form of referrals or legal requests. While there is not always a clear-cut distinction between the two categories, most referrals or legal requests fall within the parameters contained in the first two items below. Content-sharing services may also have special reporting channels or escalation pathways for specific individuals, entities, types of requests, TVEC, TVEC-related online activity or situations, such as a real-world terrorist or violent extremist event with direct online implications. The channels or pathways described below may differ or overlap slightly, as they are impacted by how companies design their respective reporting procedures.
    a. Government legal requests – Government legal requests direct a content-sharing service to remove TVEC or TVEC-related online activity that violates the law in a national or regional jurisdiction. These requests may take a number of forms, including notices and orders, and may be founded in various types of laws and legal systems. These requests may come from public authorities, like government institutions, regulators or other administrative bodies, law enforcement, or national courts.
    b. Government referrals -- Government referrals are requests by a government institution or public authority to a content-sharing service to review TVEC or TVEC-related online activity on the basis that it may violate the company's community guidelines, terms of service or other relevant guidance documents. The TVEC or TVEC-related online activity may or may not violate local law, as well.
    c. Internet Referral Units – Specialised public authorities typically housed within law enforcement bodies, with responsibility for making referrals to content-sharing services. IRUs operate

within the confines of their mandate and flag TVEC or TVEC-related online activity that violates a given country's terrorism legislation but which is referred to a company for review against the company's terms of service.

d. Online community reports – Online community reports or flags are a common mechanism for users to report TVEC or TVEC-related online activity to a content-sharing service.
e. Real-world terrorist or violent extremist event with direct online implications – A real-world terrorist or violent extremist event with direct online implications is a concurrent online manifestation of a real-world terrorist or violent extremist incident. It involves TVEC produced by a perpetrator or accomplice that appears to depict ideologically-driven murder (including attempts), torture or serious physical harm and appears to have been designed, produced and disseminated for virality – or has achieved actual virality – being shared online in a manner that presents a threat of unusually high impact (i.e. geographical / cross-platform scale), is likely to cause significant harm to communities, and therefore warrants a rapid, coordinated and decisive response by industry and relevant government agencies. For example, the live-streaming of the Christchurch attack was considered a real-world terrorist or violent extremist event with direct online implications requiring rapid response and action from industry and relevant government agencies.
f. Trusted notifiers – Some content-sharing service designate trusted notifiers or partners who are deemed particularly trustworthy, effective or are subject matter experts in a particular violation or harms type for notifying a content-sharing service of TVEC or TVEC-related online activity that violates its guidelines. Trusted notifier status may include special privileges, for example reports being prioritised, enhanced reporting functionality and increased engagement with the content-sharing service about moderation decisions. Depending on the content-sharing service, trusted notifiers may be comprised of individuals, organisations and/or government institutions.

3. Manual detection – Manual detection (also known as human detection) occurs when people manually identify user-generated TVEC or TVEC-related online activity based on a content-sharing service's guidelines and any relevant internal resources and processes, including quality control. Depending on the circumstances, these people may be employed, contracted or appointed for this purpose.

4. Automated detection – Automated detection occurs when technological tools are used in an automatic capacity, in a repeatable manner and without human triggering, to identify, surface, triage and/or action TVEC or TVEC-related online activity that violates a content-sharing service's guidelines.

Moderation – Moderation is the process of reviewing/assessing TVEC or TVEC-related online activity and deciding a course of action based on a content-sharing service's guidelines. Moderation and human review processes may be triggered by internal processes of investigations, routine checks, or from an automated triage system. They may also be triggered by external third party entity reporting or making a company aware of TVEC or TVEC-related online activity that might violate company guidelines.

1. Internal moderation – Internal moderation occurs when TVEC or TVEC-related online activity is reviewed/assessed by internal moderation teams or administrators, or by external bodies or moderation services, contracted by or at the direction of a content-sharing service to decide how to apply the company's guidelines.

2. User moderation – User moderation, or community-based moderation, occurs when a content-sharing service's users or community moderate TVEC or TVEC-related online activity directly on the service. This may occur through a removal system or a voting system which allows users to register approval or disapproval.

3. Automated moderation – Automated moderation occurs when technological tools are used automatically, in a repeatable manner to action identified TVEC or TVEC-related online activity that violates company guidelines.

4. Manual moderation – Manual moderation (also known as human moderation) occurs when people manually review/assess user-generated TVEC or TVEC-related online activity based on the company guidelines, any relevant internal resources and processes and, in some cases, the subject matter expertise or socio-linguistic understanding of the moderator. Depending on the circumstances, these people may be employed, contracted or appointed for this purpose.

5. Hybrid moderation system – A hybrid system is a mix of automated and manual detection and moderation. Content-sharing services most commonly use hybrid systems.

6. Activity-based moderation – Moderation decisions based on online user TVEC-related online activity rather than the specific pieces of content a user shares. In essence, this means that content shared by users and/or user accounts might be actioned despite a specific piece of content not having strictly violated company policy. Such moderation can rely on methods such as, but not limited to, user typologies, accounts or access signals and environment profiling.

Disabling content – See Actioning content.

Disabling accounts – See Actioning accounts.

Down-ranking – See Actioning content.

Government legal requests – See Detection and moderation, Detection, Reactive.

Government referrals – See Detection and moderation, Detection, Reactive.

Hash – A hash is a unique identifier, often likened to a signature or a fingerprint, that can be created from a digital image or video.

Hiding – See Actioning content.

Internet Referral Unit – See Detection and moderation, Detection, Reactive.

Live-stream – To live-stream is to use a content-sharing service to record and broadcast audio-visual content of an event in real-time. The transmitted content itself is also known as a live-stream.

Moderation – See Detection and moderation, Moderation.

Notification – See Actioning content.

Online community reports – See Detection and moderation, Detection, Reactive.

Providing reasons – A content-sharing service may provide a statement of reasons (such as violating or not violating company guidelines) to the user who reported certain content, requested a review, or posted the content, as well as any other user(s) affected and/or the broader community.

Quarantining – See Actioning content.

Real-world terrorist or violent extremist event with direct online implications – See Detection and moderation, Detection, Reactive.

Removing – See Actioning content.

Restoring – Restoring and/or reversing actions taken on content or accounts.

Suspending – See Actioning accounts.

Terrorist and violent extremist content (TVEC) – Content, for the purpose of the VTRF, is any type of digital information serving as a medium for terrorist and violent extremist content, such as text, video, audio and pictures. There is no universally accepted definition of terrorism or violent extremism, and congruently, of terrorist and violent extremist content. There are a number of relevant resources available for companies to consider, to help select and explain the definitions of terrorism and violence extremism they are using. Examples include the Report of the Special Rapporteur on Terrorism from 2010 to the Human Rights Council: Ten areas of best practices in countering terrorism (Section III.F "Definitions of terrorism"); the Global Research Network on Terrorism and Technology: Paper No. 7: Terrorist Definitions and Designations Lists: What Technology Companies Need to Know; and the Global Counter Terrorism Forum's Zurich-London Recommendations on Preventing and Countering Violent Extremism and Terrorism Online.

Trusted notifiers – See Detection and moderation, Detection, Reactive.

User-generated content – Content created, uploaded or shared by a content-sharing service's users.

References

4chan. (n.d.). *'Advertise - 4chan'*. Retrieved August 31, 2019, from http://www.4chan.org/advertise

4chan. (n.d.). *Frequently Asked Questions*. Retrieved from https://www.4channel.org/faq

Ahmed, M. (2020, January 31). *After Christchurch: How Policymakers Can Respond to Online Extremism*. Retrieved from Tony Blair Institute for Global Change: https://institute.global/policy/after-christchurch-how-policymakers-can-respond-online-extremism

Alexa. (2019). *The top 500 sites on the web.* Retrieved from Alexa: https://www.alexa.com/topsites/global;0

Alexander, J. (2019, August 12). *Verizon is selling Tumblr to WordPress' owner*. Retrieved from The Verge: https://www.theverge.com/2019/8/12/20802639/tumblr-verizon-sold-wordpress-blogging-yahoo-adult-content

Amazon . (n.d.). *Amazon.com Help: Law Enforcement Information Requests*. Retrieved from Amazon: https://www.amazon.com/gp/help/customer/display.html?nodeId=GYSDRGWQ2C2CRYEF

Apple. (n.d.). *Privacy - About Apple's Transparency Report*. Retrieved from Apple: https://www.apple.com/legal/transparency/about.html

Arthur, R. (2019, July 10). *We Analyzed More Than 1 Million Comments on 4chan. Hate Speech There Has Spiked by 40% Since 2015.* Retrieved from Vice: https://www.vice.com/en_us/article/d3nbzy/we-analyzed-more-than-1-million-comments-on-4chan-hate-speech-there-has-spiked-by-40-since-2015

Australian Government, F. R. (2019). *Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019.* Retrieved from https://www.legislation.gov.au/Details/C2019A00038

Automattic. (n.d.). *Transparency Report*. Retrieved from Automattic: https://transparency.automattic.com/

Barnes, L. (2019, January 17). *One month after controversial adult-content purge, far-right pages are thriving on Tumblr*. Retrieved from Think Progress: https://thinkprogress.org/far-right-content-survived-tumblr-purge-36635e6aba4b/

Barret, P. M. (2020). *Regulating Social Media: The Fight overn Section 230 - and Beyond.* NYU / STERN - Center for Business and Human Rights.

Bennett, C. a. (2019). *Extremism, George Washington University.* Retrieved from https://extremism.gwu.edu/sites/g/files/zaxdzs2191/f/EncryptedExtremism.pdf

Bicknell, Z. (2018, September 27). *What Video Platform Should I Use?* Retrieved from The UK Domain: https://www.theukdomain.uk/what-video-platform-should-i-use/

British Broadcasting Corporation (BBC). (2019, October 10). *Germany shooting: 2,200 people watched on Twitch*. Retrieved from BBC: https://www.bbc.com/news/technology-49998284

Carmen, A. (2015, December 9). *Filtered extremism: how ISIS supporters use Instagram*. Retrieved from The Verge: https://www.theverge.com/2015/12/9/9879308/isis-instagram-islamic-state-social-media

Cheah, M. (2019, June 26). *Important updates to our content guidelines - Vimeo Blog*. Retrieved from Vimeo: https://vimeo.com/blog/post/important-updates-to-our-content-guidelines/

Chen, W. (2020, April 1). *The top Chinese short-video apps in 2020 vying to grab your attention with fast content*. Retrieved from KrASIA: https://kr-asia.com/the-top-chinese-short-video-apps-in-2020-vying-to-grab-your-attention-with-fast-content

Christchurch Call . (2019). *Christchurch Call.* Retrieved from https://www.christchurchcall.com/call.html

Clicky, S. (2017, December 6). *Tackling Extremist Content on WordPress.com*. Retrieved from Transparency Report: https://transparency.automattic.com/2017/12/06/tackling-extremist-content-on-wordpress-com/

Clifford, B., & Powell, H. (2019). *Encrypted Extremism - Inside the English-Speaking Islamic State Ecosystem on Telegram.* The George Washington University, Program on Extremism.

Counter Extremism Project. (2018, August 17). *On Anniversary Of Barcelona Attacks, ISIS Continues Its Expansion*. Retrieved from Counter Extremism Project: https://www.counterextremism.com/press/anniversary-barcelona-attacks-isis-continues-its-expansion

Counter Terrorism Project. (n.d.). *Extremists & Online Propaganda.* Retrieved from Counter Terrorism Project: https://www.counterextremism.com/extremists-online-propaganda

Cox, J. (2019, April 19). *36 Days After Christchurch, Terrorist Attack Videos Are Still on Facebook*. Retrieved from Vice: https://www.vice.com/en_us/article/43jdbj/christchurch-attack-videos-still-on-facebook-instagram

Cuthbertson, A. (2019, December 03). *TikTok secretly loaded with Chinese surveillance software, lawsuit claims*. Retrieved from Independent: https://www.independent.co.uk/life-style/gadgets-and-tech/news/tiktok-china-data-privacy-lawsuit-bytedance-a9230426.html

Datanyze. (2020, September). *Market Share / File Sharing.* Retrieved from Datanyze: https://www.datanyze.com/market-share/file-sharing--198/Datanyze%20Universe

Daum Kakao. (n.d.). *Transparency Report, Kakao Privacy Policy*. Retrieved from Kakao: http://privacy.daumkakao.com/en/transparence/report/request

DCMS. (2020, February 12). *Online Harms White Paper - Initial consultation response*. Retrieved from https://www.gov.uk/government/consultations/online-harms-white-paper/public-feedback/online-harms-white-paper-initial-consultation-response

Dearden, L. (2019, August 9). *Far-right extremists 'encouraged copycat terror attacks' after Christchurch mosque shootings.* Retrieved from The Independent: https://www.independent.co.uk/news/uk/crime/far-right-terror-plots-uk-muslims-christchurch-attack-white-a9050511.html

Department of the Prime Minister and Cabinet, A. (2019, June 21). *Australian Taskforce to Combat Terrorist and Extreme Violent Material Online.* Retrieved June 5, 2019, from https://www.pmc.gov.au/sites/default/files/publications/combat-terrorism-extreme-violent-material-online.pdf

DeviantArt Media Kit. (n.d.). *There's No Place Like DeviantArt*. Retrieved from https://deviantartads.com/

DeviantArt. (n.d.). *What happens when my account is banned?* Retrieved from DeviantArt: https://www.deviantartsupport.com/en/article/what-happens-when-my-account-is-

banned

DeviantArt. (n.d.). *What is your policy around account suspensions?* Retrieved from DeviantArt: https://www.deviantartsupport.com/en/article/what-is-your-policy-around-account-suspensions

DeviantArt. (n.d.). *What policy guidelines are there on comments, Journals, statuses, and general interactions?* Retrieved from DeviantArt: https://www.deviantartsupport.com/en/article/what-policy-guidelines-are-there-on-comments-journals-statuses-and-general-interactions

Dilger, D. E. (2015, November 21). *Another security manual recommends using Apple iMessage: this time, ISIS* . Retrieved from appleinsider: https://appleinsider.com/articles/15/11/21/another-security-manual-recommends-using-apple-imessage-this-time-isis-

Discord. (2019). *Discord Transparency Report: Jan 1 — April 1* . Retrieved from Discord Blog: https://blog.discordapp.com/discord-transparency-report-jan-1-april-1-4f288bf952c9?gi=e7efc9d05321

Discord. (2020). *Discord Transparency Report: April — Dec 2019*. Retrieved from https://blog.discord.com/discord-transparency-report-april-dec-2019-7e6d43a9bcb8

Dropbox. (n.d.). *Transparency Overview*. Retrieved from Dropbox: https://www.dropbox.com/en_GB/transparency

Dropbox. (n.d.). *Who can see the stuff in my Dropbox account? Dropbox Help*. Retrieved from Dropbox: https://help.dropbox.com/accounts-billing/security/file-access

Duarte, N., Llanso, E., & Loup, A. (2017). *Mixed Messages? The Limits of Automated Social Media Cotent Analysis.* Center for Democracy & Technology.

EDRi. (2019, October 17). *Trilogues on terrorist content: Upload or re-upload filters? Eachy peachy*. Retrieved from EDRi: https://edri.org/our-work/trilogues-on-terrorist-content-upload-or-re-upload-filters-eachy-peachy/

Electronic Frontier Foundation. (2020, October). *Urgent: EARN IT Act Introduced in House of Representatives*. Retrieved from https://www.eff.org/deeplinks/2020/10/urgent-earn-it-act-introduced-house-representatives

Elmer-Dewitt, P. (2019, January 17). *Information: Facebook's Messenger has overtaken Apple's iMessage.* Retrieved from 247wallstreet.com: https://247wallst.com/technology-3/2019/01/17/apple-facebook-messaging/

European Commission. (2020). *Proposal for a regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC.*

European Commission. (2020). *Report from the Commission to the European Parliament and the Council based on Article 29(1) of Directive (EU) 2017/541 on combating terrorism and replacing Council Framework Decision 2002/475/JHA and amending Council Decision 2005/671/JHA.*

European Commission. (2020, June 02). *The Digital Services Act package*. Retrieved from https://ec.europa.eu/digital-single-market/en/digital-services-act-package

Facebook. (2017-2020). *Community Standards Enforcement Report - Dangerous Organisations: Terrorism and Organised Hate.* Retrieved from Facebook: https://transparency.facebook.com/community-standards-enforcement#terrorist-propaganda

Facebook. (2018, November 8). *Hard Questions: What Are We Doing to Stay Ahead of Terrorists? .* Retrieved from Facebook: https://about.fb.com/news/2018/11/staying-

ahead-of-terrorists/

Facebook. (2019, September 17). *Combating Hate and Extremism*. Retrieved from Facebook: https://about.fb.com/news/2019/09/combating-hate-and-extremism/

Facebook. (2019, May 23). *Measuring Prevalence of Violating Content on Facebook*. Retrieved from Facebook: https://about.fb.com/news/2019/05/measuring-prevalence/

Facebook. (2020, May 12). *An Update on Combating Hate and Dangerous Organizations*. Retrieved from Facebook: https://about.fb.com/news/2020/05/combating-hate-and-dangerous-organizations/

Facebook. (2020). *Community Standards Enforcement Report: Dangerous Organizations*. Retrieved June 5, 2020, from https://transparency.facebook.com/community-standards-enforcement#dangerous-organizations

Facebook. (n.d.). *Community Standards, 1. Violence and Incitement.* Retrieved from Facebook Web site: https://www.facebook.com/communitystandards/credible_violence

Facebook. (n.d.). *Community Standards, 2. Dangerous Individuals and Organizations.* Retrieved from Facebook Web site: https://www.facebook.com/communitystandards/dangerous_individuals_organizations

Facebook. (n.d.). *Understanding the Community Standards Enforcement Report .* Retrieved from Facebook Transparency: https://transparency.facebook.com/community-standards-enforcement/guide

Fisher-Birch, J. (2018, March 13). *Terror on Tumblr*. Retrieved from Counter Terrorism Project: https://www.counterextremism.com/blog/terror-tumblr

Frier, S. (2018, April 4). *Facebook Scans the Photos and Links You Send on Messenger.* Retrieved from Bloomberg: https://www.bloomberg.com/news/articles/2018-04-04/facebook-scans-what-you-send-to-other-people-on-messenger-app

G20. (2017). *The Hamburg G20 Leaders' Statement on Countering Terrorism*. Retrieved from https://www.mofa.go.jp/files/000271330.pdf

G20. (2019). *G20 Osaka Leaders' Statement on Preventing Exploitation of the Internet for Terrorism and Violent Extremism Conducive to Terrorism (VECT)*. Retrieved from Digital Watch Observatory: https://dig.watch/instruments/g20-osaka-leaders-statement-preventing-exploitation-internet-terrorism-and-violent

G7. (2019). *G7 Digital Ministers Chair's Summary*. Retrieved from https://www.economie.gouv.fr/files/files/2019/G7/G7Num/Chairs_summary_version_finale_ENG.pdf

GIFCT. (2020). *GIFCT Transparency Report - July 2020*. Retrieved from https://gifct.org/transparency/

GIFCT. (2021). *Membership*. Retrieved from https://gifct.org/membership/

GIFCT. (n.d.). *Global Internet Forum to Counter Terrorism: Evolving an Institution*. Retrieved from https://gifct.org/about/

GIFCT. (n.d.). *Join Tech Innovation*. Retrieved from https://gifct.org/joint-tech-innovation/

Google. (2010-2020, January-May). *Government requests to remove content - Google Transparency Report*. Retrieved from Google: https://transparencyreport.google.com/government-removals/overview?hl=en_GB

Google. (2019, January 22). *Google Drive Terms of Service*. Retrieved from Google: https://www.google.com/drive/terms-of-service/

Google. (n.d.). *Abuse program policies and enforcement - Docs Editors Help*. Retrieved from

Google: https://support.google.com/docs/answer/148505?visit_id=637064013896463652-1393240150&rd=1

Google. (n.d.). *Google Transparency Report*. Retrieved from https://transparencyreport.google.com/?hl=en_GB

Google. (n.d.). *Google Transparency Report*. Retrieved from https://transparencyreport.google.com/?hl=en_GB

Google. (n.d.). *Report a violation - Docs Editors Help*. Retrieved from Google: https://support.google.com/docs/answer/2463296?hl=en&ref_topic=1360897

Google. (n.d.). *Request a review of a violation - Docs Editors Help*. Retrieved from Google: https://support.google.com/docs/answer/2463328?hl=en&ref_topic=1360897

Google, Youtube. (2017-2020). *Google Transparency Report - Flags*. Retrieved from Google transparency Report: https://transparencyreport.google.com/youtube-policy/flags?request_examples=year:;flagging_reason:7;flagger_type:&lu=request_examples

Google, Youtube. (2019, June 5). *Our ongoing work to tackle hate*. Retrieved from YouTube blog: https://blog.youtube/news-and-events/our-ongoing-work-to-tackle-hate

Google, Youtube. (2020). *Appeal Community Guidelines actions.* Retrieved from Google, Youtube: https://support.google.com/youtube/answer/185111?hl=en

Google, Youtube. (2020). *Disable or enable Restricted/Safe Mode.* Retrieved from Google, Youtube: https://support.google.com/youtube/answer/174084?hl=en

Google, Youtube. (2020). *Report inappropriate content.* Retrieved from Google, Youtube: https://support.google.com/youtube/answer/2802027?hl=en

Google, Youtube. (2020). *YouTube Trusted Flagger program.* Retrieved from Google, Youtube: https://support.google.com/youtube/answer/7554338?&ref_topic=2803138

Google, YouTube. (n.d.). *Community Guidelines strike basics - YouTube Help*. Retrieved from Google, YouTube: https://support.google.com/youtube/answer/2802032

Google, YouTube. (n.d.). *Limited features for certain videos - YouTube Help*. Retrieved from Google, YouTube: https://support.google.com/youtube/answer/7458465

Google, Youtube. (n.d.). *YouTube Community Guidelines enforcement - Hate Speech*. Retrieved from Google: https://transparencyreport.google.com/youtube-policy/featured-policies/hate-speech?hl=en_GB

Google, YouTube. (n.d.). *YouTube Community Guidelines enforcement - Violent Extremism*. Retrieved from Google, YouTube: https://transparencyreport.google.com/youtube-policy/featured-policies/violent-extremism?hl=en_GB&policy_removals=period:Y2019Q2&lu=policy_removals

Google/Youtube. (2020). *Violent Criminal Organizations.* Retrieved from Google/Youtube Help: https://support.google.com/youtube/answer/9229472?hl=en&ref_topic=9282436

Government of Canada. (2021). *Regulation of social media platforms*. Retrieved from https://search.open.canada.ca/en/qp/id/pch,PCH-2020-QP-00084?wbdisable=true

Grüll, P. (2020, June 19). *German online hate speech reform criticised for allowing 'backdoor' data collection*. Retrieved from Euractiv: https://www.euractiv.com/section/data-protection/news/german-online-hate-speech-reform-criticised-for-allowing-backdoor-data-collection/

Harwell, D., & Romm, T. (2019, September 15). *TikTok's Beijing roots fuel censorship suspicion as it builds a huge U.S. audience*. Retrieved from The Washington Post:

https://www.washingtonpost.com/technology/2019/09/15/tiktoks-beijing-roots-fuel-censorship-suspicion-it-builds-huge-us-audience/

Hatmaker, T. (2019). *This led to Reddit administrators banning the entire community in question from the site.* Retrieved from The Tech Crunch: https://techcrunch.com/2019/03/15/reddit-watchpeopledie-subreddit-gore/

Hayden, M. E. (2019, June 27). *Far-Right Extremists Are Calling for Terrorism on the Messaging App Telegram*. Retrieved from Southern Poverty Law Center: https://www.splcenter.org/hatewatch/2019/06/27/far-right-extremists-are-calling-terrorism-messaging-app-telegram

Hayden, M. E. (2019, May 21). *Mysterious Neo-Nazi Advocated Terrorism for Six Years Before Disappearance*. Retrieved from Southern Poverty Law Center: https://www.splcenter.org/hatewatch/2019/05/21/mysterious-neo-nazi-advocated-terrorism-six-years-disappearance

Hern, A. (2019, September 25). *Revealed: how TikTok censors videos that do not please Beijing*. Retrieved from The Guardian: https://www.theguardian.com/technology/2019/sep/25/revealed-how-tiktok-censors-videos-that-do-not-please-beijing

HM Government. (2019, April). *Online Harms White Paper.* Retrieved June 4, 2019, from assets.publishing.service.gov.uk: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf

Huang, F. (2018, November 27). *China's Most Popular App Is Full of Hate*. Retrieved from Foreign Policy: https://foreignpolicy.com/2018/11/27/chinas-most-popular-app-is-full-of-hate/

Hymas, C. (2019, May 11). *Isil extremists using Instagram to promote jihad and incite support for terror attacks on the West* . Retrieved from The Telegraph: https://www.telegraph.co.uk/news/2019/05/11/isil-extremists-using-instagram-promote-jihad-incite-support/

Instagram. (2019, July 18). *Changes to Our Account Disable Policy*. Retrieved from Instagram: https://instagram-press.com/blog/2019/07/18/changes-to-our-account-disable-policy/

Iqbal, M. (2020, July 23). *Twitch Revenue and Usage Statistics (2020).* Retrieved from Business of Apps: https://www.businessofapps.com/data/twitch-statistics/

ISDGlobal. (n.d.). *Powering solutions to extremism and polarisation*. Retrieved from ISD Global: https://www.isdglobal.org/

Kallas, P. (2020, April 9). *Top 15 Most Popular Social Networking Sites and Apps [2020] @ Dreamgrow.* Retrieved from dreamgrow.com: https://www.dreamgrow.com/top-15-most-popular-social-networking-sites/

Katz, R. (2018, October 10). *To Curb Terrorist Propaganda Online, Look to YouTube. No, Really.* Retrieved from Wired: https://www.wired.com/story/to-curb-terrorist-propaganda-online-look-to-youtube-no-really/

Katz, R. (2019, September 1). *A Growing Frontier for Terrorist Groups: Unsuspecting Chat Apps*. Retrieved from Wired: https://www.wired.com/story/terrorist-groups-prey-on-unsuspecting-chat-apps/

Kemp, S. (2019, January 31). *Digital 2019: Global Digitial Overview*. Retrieved from https://datareportal.com/reports/digital-2019-global-digital-overview

Kemp, S. (2019, January 31). *Digital 2019: Q3 Global Digital Statshot.* Retrieved from

datareportal.com: https://datareportal.com/reports/digital-2019-q3-global-digital-statshot

Kemp, S. (2020, July 21). *More than Half of the People on Earth now Use Social Media.* Retrieved from https://datareportal.com/reports/more-than-half-the-world-now-uses-social-media

Kenny, K. (2019, April 30). *How can upcoming social media efforts be 'global' if they ignore Asia?* Retrieved from Stuff.co.nz: https://www.stuff.co.nz/national/christchurch-shooting/112284082/how-can-upcoming-social-media-efforts-be-global-if-they-ignore-asia

Kenyon, M. (2020, May 7). *WeChat Surveillance Explained*. Retrieved from The Citizen Lab : https://citizenlab.ca/2020/05/wechat-surveillance-explained/

Kinsta. (2011-2019). *Wordpress Market Share Statistics (2011-2019).* Retrieved from Kinsta: https://kinsta.com/wordpress-market-share/

Kitsune, L. (2017, October 11). *New Notifications and Reporting Updates by Lauren Kitsune on DeviantArt*. Retrieved from https://www.deviantart.com/laurenkitsune/journal/New-Notifications-and-Reporting-Updates-706864447

Knockel, J. L.-N. (2018, August 14). *(Can't) Picture This, An Analysis of Image Filtering on WeChat Moments.* Retrieved from The Citizen Lab: https://citizenlab.ca/2018/08/cant-picture-this-an-analysis-of-image-filtering-on-wechat-moments/

Knockel, J., Parsons, C., Ruan, L., Xiong, R., Crandall, J., & Deibert, a. R. (2020, May 7). *We Chat, They Watch How International Users Unwittingly Build up WeChat's Chinese Censorship Apparatus*. Retrieved from Citizen Lab: https://citizenlab.ca/2020/05/we-chat-they-watch/

Knockell, J. M.-N. (2015). Every Rose Has Its Thorn: Censorship and Surveillance on Social VideoPlatforms in China. Retrieved from https://www.usenix.org/system/files/conference/foci15/foci15-paper-knockel.pdf

Kock, R. (2020, July 23). *TikTok and the privacy perils of China's first international social media platform*. Retrieved from ProtonMail: https://protonmail.com/blog/tiktok-privacy/?utm_campaign=ww-en-2a-generic-coms_soc-social_organic&utm_content=&utm_medium=soc&utm_source=twitter&utm_term=1595520917

Lange, D. (2017, May 22). *Quora's Tolerance Of Terror Support*. Retrieved from Israellycool.com: https://www.israellycool.com/2017/05/22/quoras-tolerance-of-terror-support/

Le Monde. (2021, January 18). *Haine en ligne : des obligations de transparence pour les réseaux sociaux*. Retrieved from https://www.lemonde.fr/politique/article/2021/01/18/haine-en-ligne-des-obligations-de-transparence-pour-les-reseaux-sociaux_6066656_823448.html

Liao, S. (2018, February 28). *Discord shuts down more neo-Nazi, alt-right servers*. Retrieved from The Verge: https://www.theverge.com/2018/2/28/17062554/discord-alt-right-neo-nazi-white-supremacy-atomwaffen

LINE. (2019-2020). *LINE Content Moderation Report*. Retrieved from https://linecorp.com/en/security/moderation/2019h1

LINE. (n.d.). *Help Center*. Retrieved from Line: https://help.line.me/line/android/categoryId/20000132/3/pc?lang=en

Lix Xan Wong, K., & Shields Dobson, A. (2019). We're just data: Exploring China's social credit system in relation to digital platform ratings cultures in Westernised

democracies. *Global Media and China, 4*(2), 220-232.

Lokot, T. (2014, September 12). *Vkontakte, a Russian social network, is hosting ISIS accounts that were kicked off of Facebook and Twitter*. (M. J. Rosenthal, Editor) Retrieved from PRI: https://www.pri.org/stories/2014-09-12/isis-internet-army-has-found-safe-haven-russian-social-networks-now

Manileve, V. (2016, July15). *The Problem With Snapchat's Coverage of the Terror in Nice*. Retrieved from Slate: https://slate.com/technology/2016/07/did-snapchat-show-its-users-too-much-from-the-tragedy-in-nice.html

Marketing Land. (2018, September 18). *Quora Introduces Broad Targeting, Says Audience Hits 300 Million Monthly Users.* Retrieved from marketingland.com: https://marketingland.com/quora-introduces-broad-targeting-says-audience-hits-300-million-monthly-users-248261

Marketing to China. (2020, March 21). *Top 10 Chinese Social Media for Marketing (updated 2020)*. Retrieved from https://www.marketingtochina.com/top-10-social-media-in-china-for-marketing/

Marshall, C. (2019, May 28). *Twitch suspends streaming for new users as it fights off Artifact trolls*. Retrieved from Twitch: https://www.polygon.com/2019/5/28/18643198/twitch-artifact-section-stream-suspended

Medium. (2015, January 5). *Medium's Transparency Report (2014)*. Retrieved from Medium: https://medium.com/transparency-report/mediums-transparency-report-438fe06936ff

Medium. (n.d.). *Controversial, Suspect and Extreme Content, Medium Help Center*. Retrieved from Medium: https://help.medium.com/hc/en-us/articles/360018182453

Meetup. (2017, April 24). *Introducing Meetup's Inaugural Transparency Report*. Retrieved from The Meetup Blog: http://blog.meetup.com/inaugural-transparency-report/

Meetup. (2019, June 1). *Terms of Service* . Retrieved from Meetup: https://help.meetup.com/hc/en-us/articles/360027447252-Terms-of-Service

Microsoft. (2016, May 20). *Microsoft's approach to terrorist content online, Microsoft on the Issues*. Retrieved from Microsoft: https://blogs.microsoft.com/on-the-issues/2016/05/20/microsofts-approach-terrorist-content-online/#sm.000del1ea19zbe4duja1ve96fcc1l

Microsoft. (2019, January to June). *Contents Removals Request Report, Microsoft CSR.* Retrieved from Microsoft: https://www.microsoft.com/en-us/corporate-responsibility/content-removal-requests-report

Microsoft. (2021). *Digital Safety Content Report*. Retrieved from https://www.microsoft.com/en-us/corporate-responsibility/digital-safety-content-report?activetab=pivot_1:primaryr4

Miller, J. (2014, June 25). *Can Iraqi militants be kept off social media sites?* Retrieved from BBC News.

Odnoklassniki. (n.d.). *Help Centre*. Retrieved from https://ok.ru/help/54/367

OECD. (2020). Current approaches to terrorist and violent extremist content among the global top 50 online content-sharing services. In *OECD Digital Economy Papers.* OECD Publishing, Paris. doi:https://dx.doi.org/10.1787/68058b95-en

Patriquin, M. (2021, February 1). *With new legislation, Steven Guilbeault will make few friends in Big Tech*. Retrieved from https://financialpost.com/technology/with-new-legislation-steven-guilbeault-will-make-few-friends-in-big-tech

Penetrum Security. (n.d.). *Petrum Security Analysis of TikTok versions 10.0.8 -15.2.3.*

Perez, S. (2019, June 5). *Skype publicly launches screen sharing on iOS and Android*. Retrieved from Tech Crunch: https://techcrunch.com/2019/06/05/skype-publicly-launches-screen-sharing-on-ios-and-android/?guccounter=1

Perez, S. (2020, March 11). *TikTok to open a 'Transparency Center' where outside experts can examine its content moderation practices*. Retrieved from Tech Crunch: https://techcrunch.com/2020/03/11/tiktok-to-open-a-transparency-center-where-outside-experts-can-examine-its-moderation-practices/

Pew Research Center. (2016, January 14). *Wikipedia at 15: Millions of readers in scores of languages.* Retrieved from https://www.pewresearch.org/fact-tank/2016/01/14/wikipedia-at-15/

Pinterest. (2014-2020, January to March). *Transparency Report - Pinterest help*. Retrieved from Pinterest: https://help.pinterest.com/en-gb/article/transparency-report

Pinterest. (n.d.). *Account suspension*. Retrieved from https://help.pinterest.com/en/article/account-suspension

Powell, B. C. (2019). *Encrypted Extremism - Inside the English-Speaking Islamic State Ecosystem on Telegram.* The George Washington University .

Quay-de la Vallee, H., & Azarmi, M. (2020, August 25). *The New EARN IT Act Still Threatens Encryption and Child Exploitation Prosecutions*. Retrieved from https://cdt.org/insights/the-new-earn-it-act-still-threatens-encryption-and-child-exploitation-prosecutions/

Quora. (n.d.). *How does Quora Moderation make decisions about edit-blocks and bans? How does someone appeal this decision?* Retrieved from Quora: https://www.quora.com/How-does-Quora-Moderation-make-decisions-about-edit-blocks-and-bans-How-does-someone-appeal-this-decision

Reddit . (2019). *Transparency Report 2019*. Retrieved from Reddit : https://www.redditinc.com/policies/transparency-report-2019

Reddit Inc. (2017, April 17). *Moderator Guidelines for Healthy Communities*. Retrieved from Reddit: https://www.redditinc.com/policies/moderator-guidelines

Reddit Inc. (2018, January to December). *Transparency Report 2018*. Retrieved from Reddit: https://www.redditinc.com/policies/transparency-report-2018

Reddit Inc. (2020). *Transparency Report 2020*. Retrieved from https://www.redditinc.com/policies/transparency-report-2020-1

Reddit Inc. (n.d.). *AutoModerator*. Retrieved from Reddit help: https://mods.reddithelp.com/hc/en-us/articles/360002561632-AutoModerator

Reddit Inc. (n.d.). *Quarantined Subreddits*. Retrieved from Reddit Help: https://www.reddithelp.com/en/categories/rules-reporting/account-and-community-restrictions/quarantined-subreddits

Ruan, L. J.-N. (2016). *One App, Two Systems, How WeChat uses one censorship policy in China and another internationally.* Retrieved from The Citizen Lab: https://citizenlab.ca/2016/11/wechat-china-censorship-one-app-two-systems/

Ruan, L., Knockel, J., Ng, J. Q., & Crete-Nishihata, a. M. (2016). *One App, Two Systems, How WeChat uses one censorship policy in China and another internationally.* Retrieved from The Citizen Lab: https://citizenlab.ca/2016/11/wechat-china-censorship-one-app-two-systems/

Santa Clara University's High Tech Law Institute. (n.d.). *The Santa Clara Principles On Transparency and Accountability in Content Moderation*. Retrieved from santaclaraprinciples.org: https://santaclaraprinciples.org/

Singh, M. (2020, April 24). *Telegram hits 400M monthly active users.* Retrieved from Tech Crunch: https://techcrunch.com/2020/04/24/telegram-hits-400-million-monthly-active-users/

Site Intelligence Group Enterprise. (2018, December 11). *IS-linked Media Group Makes Foray onto Viber Messenger - Dark Web and Cyber Security*. Retrieved from Site Intelligence Group Enterprise: https://ent.siteintelgroup.com/Dark-Web-and-Cyber-Security/is-linked-media-group-makes-foray-onto-viber-messenger.html

Sky News. (2020, May 19). *FBI unlocks terrorist's iPhones and finds al Qaeda links - 'no thanks to Apple'*. Retrieved from Sky News: https://news.sky.com/story/fbi-unlocks-terrorists-iphones-and-finds-al-qaeda-links-no-thanks-to-apple-11990818

Snap Inc. (2015-2020). *Privacy Centre, Transparency Report - (1 January - 30 June 2019)*. Retrieved from Snap Inc.: https://www.snap.com/en-GB/privacy/transparency

Snap Inc. (n.d.). *Safety Centre, Report a safety concern*. Retrieved from Snap Inc.: https://www.snap.com/en-GB/safety/safety-reporting/

Solsman, J. E. (2018, July 14). *'Smule May Be the Biggest Music App You Haven't Heard Of'.* Retrieved from CNET: https://www.cnet.com/news/smule-is-the-biggest-music-app-you-never-heard-of/

START (National Consortium for the Study of Terrorism and Responses to Terrorism). (2018). *The Use of Social Media by United States Extremists.* University of Maryland. Retrieved from https://www.start.umd.edu/pubs/START_PIRUS_UseOfSocialMediaByUSExtremists_ResearchBrief_July2018.pdf

Statista. (2019, December 31). *MAU of iQiyi's mobile app in China 2016-2019 Published by Lai Lin Thomala, Jun 3, 2020 In 2019, the video app of iQiyi reached an average of 476 million monthly active users. Founded in Beijing in 2010, Baidu's iQiyi is one of the largest online video pl*. Retrieved from https://www.statista.com/statistics/1106091/china-online-video-platform-iqiyi-mobile-app-monthly-active-user-number/

Statista. (2019, August 9). *Number of global monthly active Kakaotalk users from 1st quarter 2013 to 1st quarter 2019.* Retrieved from https://www.statista.com/statistics/278846/kakaotalk-monthly-active-users-mau/

Stokel-Walker, C. (2020, March 20). *As humans go home, Facebook and YouTube face a coronavirus crisis*. Retrieved from Wired: https://www.wired.co.uk/article/coronavirus-facts-moderators-facebook-youtube

Tardi, C. (2019, August 27). Monthly Active User (MAU) . *Investopedia*. Retrieved from https://www.investopedia.com/terms/m/monthly-active-user-mau.asp

Tech Against Terrorism. (2019, April). *Analysis: ISIS use of smaller platforms and the DWeb to share terrorist content – April 2019*. Retrieved from https://www.techagainstterrorism.org/2019/04/29/analysis-isis-use-of-smaller-platforms-and-the-dweb-to-share-terrorist-content-april-2019/

Tech Against Terrorism. (2020). *The Online Regulation Series - India*. Retrieved from https://www.techagainstterrorism.org/2020/10/09/the-online-regulation-series-india/

Tech Against Terrorism. (2020). *The Online Regulation Series - Singapore*. Retrieved from https://www.techagainstterrorism.org/2020/10/05/the-online-regulation-series-singapore/

Tech Against Terrorism. (2020). *The Online Regulation Series: The United States*. Retrieved from Tech Against Terrorism: https://www.techagainstterrorism.org/2020/10/13/the-

online-regulation-series-the-united-states/

Telegram. (n.d.). *ISIS Watch*. Retrieved from Telegram: https://telegram.me/ISISwatch

Telegram. (n.d.). *Telegram Privacy Policy*. Retrieved from Telegram: https://telegram.org/privacy

Tencent. (n.d.). *Agreement on Software License and Service of Tencent Weixin*. Retrieved from https://weixin.qq.com/cgi-bin/readtemplate?lang=en&t=weixin_agreement&s=default&cc=CN

The Hindu Business Line. (2020, February 13). *Social media users to be tracked by government under new guidelines*. Retrieved from https://www.thehindubusinessline.com/info-tech/social-media/social-media-users-to-be-tracked-by-government-under-new-guidelines-report/article30807839.ece

The International Centre for the Study of Radicalisation (ICSR). (2020). *ICSR info*. Retrieved from The International Centre for the Study of Radicalisation (ICSR): https://icsr.info/

The Santa Clara Principles. (n.d.). *The Santa Clara Principles on Transparency and Accountability in Content Moderation*. Retrieved from https://santaclaraprinciples.org

Thompson, E. (2021, January 29). *Canada not exempt from social media forces that created U.S. Capitol riot, heritage minister says*. Retrieved from https://www.cbc.ca/news/politics/facebook-twitter-canada-regulation-1.5894301

Thune, J. (2020, July 29). *Thune: PACT Act Would Increase Internet Accountability and Consumer Transparency*. Retrieved from https://www.thune.senate.gov/public/index.cfm/2020/7/thune-pact-act-would-increase-internet-accountability-and-consumer-transparency

TikTok. (2019-2020). *TikTok Transparency Report*. Retrieved from https://www.tiktok.com/safety/resources/transparency-report?lang=en

Titcomb, J. (2017, November 1). *Why Google is reading your Docs* . Retrieved from The Telegraph: https://www.telegraph.co.uk/technology/2017/11/01/google-reading-docs/

Tumblr. (2019). *Tumblr Government Transparency Report*. Retrieved from Tumblr: https://static.tumblr.com/elwkrsl/u9Cqct0vc/government_transparency_report_2019.pdf

Twitch. (2020). *Transparency Report 2020*. Retrieved from https://www.twitch.tv/p/en/legal/transparency-report/

Twitch. (n.d.). *How to Use AutoMod*. Retrieved from Twitch TV: https://help.twitch.tv/s/article/how-to-use-automod?language=en_US

Twitter. (2012-2020). *Twittter Rules enforcement*. Retrieved from Twitter Transparency Reportt: https://transparency.twitter.com/en/twitter-rules-enforcement.html

Twitter. (n.d.). *Our approach to policy development and enforcement philosophy*. Retrieved from Twitter: https://help.twitter.com/en/rules-and-policies/enforcement-philosophy

Twitter. (n.d.). *Our range of enforcement options*. Retrieved from help.twitter.com: https://help.twitter.com/en/rules-and-policies/enforcement-options

United Nations Office on Drugs and Crime. (2012). *The use of the Internet for terrorist purposes.* United Nations. Vienna: United Nations Office at Vienna.

United Nations Security Council. (n.d.). *United Nations Security Council Consolidated List.* Retrieved from United Nations Security Council: https://www.un.org/securitycouncil/content/un-sc-consolidated-list

US Treasury. (2020, January 23). *OFFICE OF FOREIGN ASSETS CONTROL - Specially Designated Nationals and Blocked Persons List.* Retrieved from Treasury:

https://www.treasury.gov/ofac/downloads/sdnlist.pdf

Verizon Media. (2019). *Transparency Report*. Retrieved from Verizon Media: https://www.verizonmedia.com/transparency/index.html?guce_referrer=aHR0cHM6Ly90cmFuc3BhcmVuY3kub2F0aC5jb20vaW5kZXguaHRtbD9ndWNlX3JlZmVycmVyPWFIUjBjSE02THk5M2QzY3VkSFZ0WW14eUxtTnZiUzgmZ3VjZV9yZWZlcnJlcl9zaWc9QVFBQUFKazduZ3VNWS04dHhtNG9hWFM3TUlkNkxlUWxkMEZ5

Vimeo . (n.d.). *How does Vimeo deal with violent content? - Help Center*. Retrieved from Vimeo Zendesk: https://vimeo.zendesk.com/hc/en-us/articles/224822427-How-does-Vimeo-deal-with-violent-content-

VK. (2020). *Platform Standards*. Retrieved from https://m.vk.com/safety?lang=en&section=standarts

VK. (2020). *Safety Guidelines*. Retrieved from VK : https://m.vk.com/safety?section=social&lang=en

Wang, Z. (2017). Systematic Government Access to Private-Sector Data in China. In F. H. Dempsey (Ed.), *Bulk Collection - Systematic Government Access to Private-Sector Data.* Oxford University Press.

Weimann, G. (2014). *New Terrorism and New Media.* Commons Lab of the Woodrow Wilson International Center for Scholars. Retrieved from https://www.wilsoncenter.org/sites/default/files/new_terrorism_v3_1.pdf

Wickey, W. (2018, August 23). *Should You Use Medium As Your Business Blog Platform? [2019 Update].* Retrieved from Medium: https://medium.com/crowdbotics/medium-business-blog-platform-b8b8faa2d430

Wikimedia Foundation. (2019, June 7). *Terms of Use - Wikimedia Foundation Governance Wiki*. Retrieved from Wikimedia Foundation: https://foundation.wikimedia.org/wiki/Terms_of_Use/en

Wikimedia Foundation. (n.d.). *Transparency report*. Retrieved from Wikimedia Foundation: https://transparency.wikimedia.org/

Wikipedia. (2019, October 22). *CheckUser - Wikipedia*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Wikipedia:CheckUser

Wikipedia. (2019, December 14). *Core Content Policies - Wikipedia*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Wikipedia:Core_content_policies

Wikipedia. (2020, January 1). *Administration - Wikipedia*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Wikipedia:Administration#Human_and_legal_administration

Wikipedia. (2020, January 26). *Criteria for speedy deletion - Wikipedia*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Wikipedia:Criteria_for_speedy_deletion#Procedure_for_administrators

Wikipedia. (2020, January 23). *Deletion process - Wikipedia*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Wikipedia:Deletion_process

Wikipedia. (2020, January 28). *Oversight - Wikipedia*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Wikipedia:Oversight

Wikipedia. (2020, January 27). *What Wikipedia is not - Wikipedia*. Retrieved from Wikipedia.

Word Press. (n.d.). *Terrorist Activity - Support - Word Press.com*. Retrieved from Word Press: https://en.support.wordpress.com/terrorist-activity/

WordPress.com. (n.d.). *Suspended Content and Sites*. Retrieved from WordPress.com:

https://en.support.wordpress.com/suspended-blogs/

Yahoo! Finance. (2019, November 13). *YY earnings surpass estimates in Q3, revenues increase*. Retrieved from Yahoo! Finance: https://finance.yahoo.com/news/yy-earnings-surpass-estimates-q3-144502223.html

Yoo, E. (2018, April 13). *Huoshan latest video platform to clean up vulgar content*. Retrieved from technode: https://technode.com/2018/04/13/huoshan-clean-up/

Youku Tudou Inc. (NYSE: YOKU). (n.d.). *Youku Tudou Inc. (NYSE: YOKU), About us - 优酷视频*. Retrieved from c.you.ku.com: https://c.youku.com/abouteg/youtu

YouTube. (2020, March 16). *Protecting our extended workforce and the community*. Retrieved from YouTube Official Blog: https://blog.youtube/news-and-events/protecting-our-extended-workforce-and

YY Inc. - IR Site. (2019, May 28). *YY Reports First Quarter 2019 Unaudited Financial Results.* Retrieved from http://ir.yy.com/news-releases/news-release-details/yy-reports-first-quarter-2019-unaudited-financial-results

Zetter, K. (2015, November 19). *Security Manual Reveals the OPSEC Advice ISIS Gives Recruits*. Retrieved from Wired: https://www.wired.com/2015/11/isis-opsec-encryption-manuals-reveal-terrorist-group-security-protocols/

Zhong, R. (2018, November 8). *At China's Internet Conference, a Darker Side of Tech Emerges*. Retrieved from The New York Times: https://www.nytimes.com/2018/11/08/technology/china-world-internet-conference.html

Notes

[1] See Section 1 of the Services' profiles in Annex B.

[2] See Sections 5 and 6 of the Services' profiles in Annex B.

[3] "MAU helps to measure an online business's general health and is the basis for calculating other website metrics. MAU is also useful when assessing the efficacy of a business's marketing campaigns and gauging both present and potential customers' experience. Investors in the social media industry pay attention when companies report MAU, as it is a [key performance indicator] that can affect a social media company's stock price" (Tardi, 2019).

[4] See the Services' profiles in Annex [B] of the first benchmarking report.

[5] Information from media outlets and other publicly available sources was used, however, in Section 10 of each profile (see Annex B), not least because the Services' governing documents rarely list concrete incidents where their technologies are exploited to further terrorist and violent extremist ends. At any rate, when used, these sources of information are duly referenced via endnotes.

[6] Facebook, YouTube, TikTok, Twitter and Google Drive.

[7] See Section 1 of the Facebook, YouTube, TikTok, Twitch, Twitter and Google Drive profiles. Arguably, Microsoft (LinkedIn, Skype and OneDrive) belongs in this group, as well, though it provides no definition of violent extremism and does not offer any examples. Similarly, Discord provides good explanations and descriptions of violent extremism and hate speech, but it does not define terrorism. Pinterest also provides good descriptions of hateful activities and content, but it does not define extremists and terrorist organisations.

[8] Instagram, Youku Tudou, iQIYI, LinkedIn, Baidu Tieba, Skype, Quora, Snapchat, Pinterest, Ask.fm, Xigua, Tumblr, Flickr, Huoshan, Haokan, Meetup, Dropbox, Microsoft OneDrive and Wordpress.com.

[9] See Section 1 of the Instagram, Youku Tudou, iQIYI, Kuaishou, LinkedIn, Baidu Tieba, Skype, Quora, Snapchat, Pinterest, Ask.fm, Xigua, Discord, Tumblr, Flickr, Huoshan, Haokan, Meetup, Dropbox, Microsoft OneDrive and Wordpress.com profiles.

[10] WeChat, Instagram, QQ, Youku Tudou, iQIYI, Douban, LinkedIn, Baidu Tieba, Vimeo, Twitch, Medium, Odnoklassniki, KaKaoTalk, Meetup and MySpace.

[11] See Section 1 of the WeChat, Instagram, QQ, Youku Tudou, iQIYI, Kuaishou, Douban, LinkedIn, Baidu Tieba, Vimeo, Medium, Odnoklassniki, and Meetup profiles.

[12] WhatsApp, iMessage/FaceTime, QZone, Weibo, Reddit, Viber, IMO, Telegram, LINE, VK, YY Live, Discord, Smule, DeviantArt, 4chan and Wikipedia.

[13] See Section 1 of the WhatsApp, iMessage/FaceTime, QZone, Weibo, Reddit, Viber, IMO, Telegram, LINE, VK, YY Live, Smule, DeviantArt, 4chan and Wikipedia profiles.

[14] See Section 1 of the Facebook and Instagram profiles.

[15] See Section 7 of the YouTube profile, and Section 1 of the Skype, Quora, Microsoft OneDrive and

Wordpress.com profiles.

[16] See Section 1 of the VK profile.

[17] See Section 1 of the WhatsApp, iMessage/Facetime, WeChat, QQ, Youku Tudou, Weibo, QZone, iQIYI, Reddit, Kuaishou, Telegram, Snapchat, Pinterest, Twitter, Douban, Baidu Tieba, Xigua, Viber, Discord, Vimeo, IMO, LINE, Huoshan, Ask.fm, YY Live, Twitch, Tumblr, Flickr, Medium, Odnoklassniki, Haokan Video, Smule, KakaoTalk, DeviantArt, Meetup, 4chan, Google Drive, Dropbox and Wikipedia profiles.

[18] Encryption keeps communications confidential between the sender and receiver, so that no third-party can access the communications, including the company providing the service. Encryption also protects information stored on computers, mobile phones, and other digital devices, ensuring that if the device is lost or stolen the information on the device is protected. Encryption allows individuals to freely express themselves, to exchange personal and other sensitive information, and to protect their data. On the other hand, malicious actors are able to abuse the confidentiality, privacy and security encryption affords to plot and coordinate terrorist attacks, engage in organised crime and preserve their anonymity. Accordingly, encryption presents a complicated trade-off between, on one hand, privacy and security, and on the other hand, law enforcement and transparency reporting.

[19] See Section 2 – Differences between Current TVEC Transparency Reports of the first benchmarking report.

[20] Facebook, YouTube, WhatsApp, Facebook Messenger, iMessage/FaceTime, Instagram, TikTok, Weibo, Reddit, Twitter, LinkedIn, Baidu Tieba, Skype, Quora, Snapchat, Viber, Pinterest, Vimeo, Telegram, LINE, Ask.fm, Xigua, Tumblr, Flickr, Houshan, VK, Medium, Odnoklassniki, Discord, Smule, KaKaoTalk, DeviantArt, Meetup, 4chan, MySpace, Google Drive, Dropbox, OneDrive, WordPress.com and Wikipedia.

[21] See Section 5 of the Facebook, YouTube, WhatsApp, Facebook Messener, iMessage/FaceTime, Instagram, TikTok, Weibo, Reddit, Kuaishou,Twitter, LinkedIn, Baidu Tieba, Skype, Quora, Snapchat, Viber, Pinterest, Vimeo, Telegram, LINE, Ask.fm, Xigua, Tumblr, Flickr, Houshan, VK, Medium, Odnoklassniki, Discord, Smule, KaKaoTalk, DeviantArt, Meetup, 4chan, Google Drive, Dropbox, OneDrive, WordPress.com and Wikipedia profiles.

[22] Reddit, Viber, Twitch, Flickr, VK, Odnoklassniki, KaKaoTalk, DeviantArt, 4chan and Wikipedia.

[23] See Section 4 and 5 of the Reddit, Viber, Twitch, Flickr, VK, Odnoklassniki, KaKaoTalk, DeviantArt, 4chan and Wikipedia profiles.

[24] The expression “at least” is included because it was not possible to determine, based on some Services’ publicly disclosed information, the kind of activities and processes they implement to enforce their ToS and other governing documents.

[25] Facebook, YouTube, WhatsApp, Facebook Messenger, WeChat, Instagram (Hash Sharing Consortium member), TikTok, Reddit (Hash Sharing Consortium member), Twitter, LinkedIn (Hash Sharing Consortium member), Skype (indirect membership of GIFCT through Microsoft), Snapchat (Hash Sharing Consortium member), Pinterest (GIFCT member), LINE, Ask.fm (Hash Sharing Consortium member), Twitch (indirect membership of GIFCT through Amazon), VK, YY Live, Google Drive, Dropbox (GIFCT member) and OneDrive (GIFCT member).

[26] Again, the expression “at least” is included because it was not possible to determine, based on some Services’ publicly disclosed information, the kind of activities and processes they implement to enforce their ToS and other governing documents. See for example Section 5 of the QQ, Youku Tudou, QZone, Weibo, iQIYI, Douban, Baidu Tieba, YY Live, Xigua, Huoshan and Haokan profiles.

[27] See Section 5 of the Facebook, YouTube, WhatsApp, Facebook Messenger, WeChat, Instagram (GIFCT member), TikTok, Reddit (Hash Sharing Consortium member), Twitter, LinkedIn (Hash Sharing Consortium member), Skype (indirect membership of GIFCT through Microsoft), Snapchat (Hash Sharing Consortium member), Pinterest (GIFCT member), Viber, Discord (GIFCT member), LINE, Ask.fm (Hash Sharing Consortium member), Twitch (indirect membership of GIFCT through Amazon), VK, YY Live, Google Drive, Dropbox (GIFCT member) and OneDrive (GIFCT member) profiles.

[28] Facebook, YouTube, Facebook Messenger, Instagram, Reddit, Twitter, Quora, Pinterest, Vimeo, Ask.fm, Twitch, Tumblr, VK, Medium, Odnoklassniki, Smule, KaKaoTalk, DeviantArt, Meetup, Dropbox and Wordpress.com.

[29] See Section 4.1 of the Facebook, YouTube, Facebook Messenger, WhatsApp, Instagram, Reddit, Snapchat, Twitter, Quora, Pinterest, Vimeo, Ask.fm, Twitch, Tumblr, VK, Medium, Odnoklassniki, Smule, KaKaoTalk, DeviantArt, Meetup, Dropbox and Wordpress.com profiles.

[30] Facebook, YouTube, WhatsApp, Facebook Messenger, Instagram, TikTok, Reddit, Twitter, Quora, Pinterest, Vimeo, LINE, Ask.fm, Twitch, Tumblr, VK, Medium, Discord, KaKaoTalk, DeviantArt, Meetup, 4chan and Wordpress.com profiles.

[31] See Section 4.2 of the Facebook, YouTube, WhatsApp, Facebook Messenger, Instagram, TikTok, Reddit, Kuaishou, Twitter, Snapchat, Quora, Viber, Pinterest, Vimeo, LINE, Ask.fm, Twitch, Tumblr, VK, Medium, Discord, KaKaoTalk, DeviantArt, Meetup, 4chan, Dropbox and Wordpress.com profiles.

[32] WhatsApp, iMessage/FaceTime, WeChat, Instagram, QQ, TikTok, Weibo, iQIYI, Douban, LinkedIn, Quora, Snapchat, Pinterest, IMO, Ask.fm, VK, Haokan, Odnoklassniki, Smule, Meetup, MySpace and OneDrive.

[33] See Section 4 and 5 of the iMessage/FaceTime, WeChat, QQ, Weibo, iQIYI, Kuaishou, Douban, LinkedIn, Quora, Pinterest, IMO, Ask.fm, VK, Haokan, Odnoklassniki, Smule and Meetup profiles. Use of the word 'may' or the expression 'reserves the right to review', in particular, are very common.

[34] See Sections 4 and 5 of the WeChat, QQ, Youku Tudou, QZone, Weibo, iQIYI, Kuaishou, Douban, Baidu Tieba, YY Live, Xigua, Huoshan and Haokan Video profiles.

[35] See for example https://extremism.gwu.edu/sites/g/files/zaxdzs2191/f/EncryptedExtremism.pdf and https://www.counterextremism.com/sites/default/files/Extremists%20and%20Online%20Propaganda_040918.pdf

[36] http://www.terrorismanalysts.com/pt/index.php/pot/article/view/607/1200

[37] See consultation process and associated documentation at https://www.communications.gov.au/have-your-say/consultation-bill-new-online-safety-act

[38] For further information on Australia's abhorrent violent material and ISP blocking schemes, please visit the following references:

- eSafety Blog on Range of Christchurch Tools & Powers: https://www.esafety.gov.au/about-us/blog/christchurch-shifted-online-world-its-axis
- eSafety AVM Fact Sheet: https://www.esafety.gov.au/sites/default/files/2020-03/eSafety-AVM-factsheet.pdf
- eSafety ISP blocking Fact Sheet: https://www.esafety.gov.au/sites/default/files/2020-03/eSafety-ISP-Blocking-factsheet.pdf
- eSafety press release on landmark ISP blocking protocol: https://www.esafety.gov.au/about-us/newsroom/blocking-viral-spread-terrorist-content-online

[39] See https://eur-lex.europa.eu/legal-content/en/TXT/?qid=1608117147218&uri=COM%3A2020%3A825%3AFIN

[40] The text of the new legislation was made available in English as part of the European notification procedure: https://ec.europa.eu/growth/tools-databases/tris/en/index.cfm/search/?trisaction=search.detail&year=2020&num=65&mLang=EN

[41] Further information including a summary of the NetzDG regulations and answers to frequently asked questions can be found in English at https://www.bmjv.de/DE/Themen/FokusThemen/NetzDG/NetzDG_EN_node.html

[42] See text at https://www.gov.ie/en/publication/d8e4c-online-safety-and-media-regulation-bill/

[43] Available at: http://www.legislation.govt.nz/bill/government/2020/0268/latest/LMS294551.html

[44] Available at: https://www.legislation.govt.nz/act/public/2015/0063/latest/DLM5711810.html

[45] This profile is about the Facebook platform itself rather than the entire company, so it does not include Messenger, Instagram or WhatsApp.

[46] The YouTube Trusted Flagger program was developed by YouTube to help provide robust tools for individuals, government agencies, and non-governmental organizations (NGOs) that are particularly effective at notifying YouTube of content that violates their Community Guidelines. https://support.google.com/youtube/answer/7554338?&ref_topic=2803138

[47] See Section 3 of the Report.

[48] It must be noted that these Terms apply only to QQ users anywhere in the world, except if they belong in any of the following categories: (a) a QQ user in the People's Republic of China; (b) a citizen of the People's Republic of China using QQ anywhere in the world; or (c) a Chinese-incorporated company using QQ anywhere in the world. Users in those categories are governed by the Terms of Service applicable to PRC users, available at https://www.qq.com/contract.shtml

[49] Qzone can be accessed outside China only through QQ International.

[50] These ToS applies to users outside China. QZone users in China are governed by the Terms of Service applicable to PRC users, available at https://www.qq.com/contract.shtml.

[51] Tumblr stated that it participates in the Hash Sharing Consortium; however, as of September 2020, the GIFCT website contains no information about this membership.