OECD publishing

TRANSPARENCY REPORTING ON TERRORIST AND VIOLENT EXTREMIST CONTENT ONLINE 2022

OECD DIGITAL ECONOMY PAPERS
October 2022 **No. 334**

OECD
BETTER POLICIES FOR BETTER LIVES

Foreword

This is the third instalment in a series of annual reports that take stock of the current policies and procedures related to terrorist and violent extremist content (TVEC) of the world's leading online platforms and other online content sharing services. The first two reports were Current Approaches to Terrorist and Violent Extremist Content among the Global Top-50 Online Content Sharing Services (2020) and Transparency Reporting on Terrorist and Violent Extremist Content Online: An update on the global top-50 content sharing services (2021). This year's edition contains a new feature: it tracks the 50 services on which the most TVEC appears, in addition to continuing to track the global top-50 most widely used services in general.

This report was written by Dr Tomas Llanos of University College London. It incorporates the oral and written feedback from delegates on earlier drafts, as well as feedback from the companies profiled in Annexes B and D. This report was approved and declassified by the Committee on Digital Economy Policy by written procedure on 21 September 2022 and prepared for publication by the OECD Secretariat.

This document, as well as any data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

Note to Delegations:

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

DSTI/CDEP(2021)21/FINAL

Table of contents

Executive Summary

This is the third edition of the OECD's annual benchmarking report examining the policies and procedures related to terrorist and violent extremist content (TVEC) of online content-sharing services, with a focus on transparency reporting. It builds on the two prior editions, provides an objective and factual snapshot in time of the services' TVEC policies and procedures, and is comprised of two main parts. The first explores the evolution of the world's top-50 online content-sharing services' ("Popular Services") approaches to TVEC online since the publication of the second edition. The second adds a new approach that analyses the TVEC-relevant policies, procedures and practices of the online content-sharing services that are used the most for disseminating TVEC ("Intensive Services").

1. Minimal overlap between the Popular and Intensive Services, and greater variance in the elaboration and detail provided in the latter group's guidelines, terms and conditions.

Only eleven services appear on both the Popular and Intensive lists (Discord, Dropbox, Facebook, Google Drive, Instagram, Telegram, TikTok, Twitter, VK, WhatsApp, YouTube). Thirteen of the Intensive Services have no prohibition on TVEC. Only seven are comprehensive and clear about their content moderation approaches and mechanisms, and seventeen provide no information at all on that subject. Overall, mainstream services with greater financial power tend to be more transparent about their TVEC-related procedures, mechanisms and initiatives than less popular and fringe services.

2. Positive trends in transparency reporting on TVEC among Popular Services, but the practice is rare among Intensive Services.

Fifteen of the Popular Services now issue transparency reports on TVEC, as compared to just five in the first edition and eleven in the second. Growth in the number of companies engaging in this practice is a positive development, as are the trends towards greater definitional clarity and convergence in some of the metrics they report. Nevertheless, significant variance in the types of information reported remains, thereby making data aggregation and cross-platform comparisons difficult.

By contrast, only eight Intensive Services issue transparency reports on TVEC and the majority of those (six) also appear in the Popular Services list. Thus, of the 39 Intensive Services that are not also Popular Services, only two issued transparency reports expressly addressing TVEC. The paucity of transparency reporting on TVEC among (non-Popular) Intensive Services may have something to do with the fact that many of them are operated by terrorist and violent extremist groups or sympathisers. Some video-sharing and news aggregator platforms like Brandnewtube, Worldtruthvideos and Mzwnews.com are specifically designed to cater to TVEC that has been removed from Popular Services, while certain other Intensive Services like Nordfront.dk, Lookaheadamerica.org, Patriotfront.us, Vastarinta.com and Nordicresistancemovement.org openly espouse hateful ideologies and white supremacist views.

3. Service size and user privacy are not insurmountable barriers to transparency reporting

One of the two Intensive Services that issue TVEC-specific transparency reports – Mega – is a privacy-driven cloud storage and chat platform offering end-to-end encryption and zero-knowledge privacy. This shows that the inability to monitor user content and communications directly, due to end-to-end encryption or other cryptographic protocols, is not a barrier to publishing transparency reports on content moderation of illegal and/or harmful content, including TVEC. The other service – Justpaste.it – is a text- and images-sharing platform with limited monetisation streams, and its staff is extremely lean (it is run by only one person). Hence, whereas the time, resources and costs involved in devising content moderation policies and procedures and producing transparency reports on content moderation must be acknowledged as challenges for small platforms, the example of this service shows that it can be done.

4. Displacement effects are shifting the landscape

Some online content-sharing services have strengthened their content moderation efforts in response to being used to spread TVEC. Actions taken by mainstream platforms to address TVEC have resulted in switching patterns amongst terrorist and violent extremist groups and individuals. These bad actors tend to find shelter in small- or micro-platforms that lack either the financial and technical resources to make their services inhospitable to TVEC or the will to do so. The latter platforms are either security-oriented, encryption-based platforms that afford high degrees of privacy, anonymity and confidentiality, or they are fringe, Alt-Tech platforms that seek to capitalise on mainstream platforms' deplatforming actions, seeing them as opportunities to expand their user base and provide infrastructure to free speech fundamentalists, provocateurs, white supremacists and other extremist groups. These developments suggest that an effective response to combat TVEC online requires broadening the focus beyond the largest, most popular platforms to the entirety of the online content-sharing ecosystem. The TVEC online problem is larger than previous editions of this report suggested because so many of the services on which TVEC is disseminated provide little or no transparency on their moderation efforts, or engage in little or no content moderation in the first place. For these reasons, this report features a new focus on the content-sharing services having the greatest impact on and interaction with the overall TVEC online landscape.

5. Growing concerns about the impact of content moderation on privacy, freedom of expression and due process

Increased content moderation by online services continues to raise concerns about undue impacts on fundamental rights such as privacy, freedom of expression and due process, as it is feared that overly intrusive monitoring of content to remove TVEC may impinge upon such rights. The COVID-19 pandemic, related lockdowns and work-from-home measures led some services to raise their reliance on automated monitoring systems to detect and remove TVEC. That had the unintended consequence of increasing the removal of lawful content and undue censorship more generally, thus heightening concerns about individuals′ability to exercise their rights. Research in this edition highlights the importance of human intervention to ensure fair and effective content moderation that protects users' fundamental rights and due process.

Introduction

Internet-based technologies such as social media, video, chat and messaging apps, streaming and file sharing have afforded new forms of participation and interaction amongst individuals and businesses across spatial barriers. News, information and a diverse range of other content can be seamlessly accessed and shared with a laptop, mobile phone and an increasing number of devices with Internet connections. That has opened multiple avenues for socio-economic development and well-being. It has become easier to gain and disseminate knowledge, convey ideas, start a business, raise capital, express opinions and interact with the government, for example (OECD, 2017[1])[1]. However, whilst the Internet has largely had positive impacts on entrepreneurship, commerce, education, freedom of expression and democratic deliberation, it has also been abused, including by terrorists and violent extremists who use it to recruit supporters, glorify their atrocities and otherwise prepare and facilitate terrorist activity (European Commission, 2018[2]).

The abuse of the Internet for terrorist and extremist purposes has risen in recent years. Research has revealed that jihadi terrorist organisations like Al-Qaeda and the Islamic State ("IS") make extensive use of social media to advertise their ideology, enrol new members, manage digital training environments and network with other terrorist organisations (Droogan, Waldek and Blackhall, 2018[3]; Weimann, 2016[4]). Jihadists are hardly the only bad actors abusing the Internet to promote terrorist or violent extremist agendas, though.

The dissemination of terrorist and violent extremist content (TVEC) online has proved to be an essential component of terrorist groups' radicalisation campaigns (European Commission, 2018[2]). All of the terrorist attacks that took place in the United Kingdom in 2017 had an online component (HM Government, April 2019[5]). Some of the Internet's virtues – i.e. ease of access to large audiences and seamless flow of information – are as advantageous to groups wishing to amplify terrorist and extremist messages as they are to legitimate users. In particular, online communications expand the reach of terrorist and extremist ideologies and consolidate communities of like-minded individuals who share extremist views, thus rendering radicalisation more likely. Terrorist and violent extremist propaganda can incite individuals to perpetrate massacres or other acts of violence autonomously, thereby motivating lone actor attacks which inspire others to commit similar actions. As a result, the threat of violent incidents such as those in Christchurch (New Zealand), Poway (USA), El Paso (USA), Baerum (Norway) and Halle (Germany), the perpetrators of which were part of similar online communities and took inspiration from one another (EUROPOL, 2020[6]), increases.

Box 1 sets forth a typology of online services with their specific TVEC use-cases.

Box 1. Typology of online services used for TVEC dissemination

- **Social media platforms**: they offer the best opportunity to reach a large external audience and to have bilateral engagements with members, supporters, and wider populations.
- **Messaging apps**: they offer an easy, secure, and often free means of both internal and external communication. Most messaging apps frequently used by terrorist actors are protected by either end-to-end or client-server encryption (or give the impression of such encryption).
- **Alt-tech platforms**: A variety of platforms have emerged in the past few years that claim to offer an alternative to larger mainstream platforms like Facebook, Twitter, and YouTube. These platforms often explicitly market themselves as "free speech" platforms, or ones that oppose the "censorship" of larger platforms. Some alt-tech platforms use blockchain-based decentralised technology. Both of these qualities are attractive to terrorists (and violent extremists) and maximise their chances of online stability. Many alt-tech platforms are Video Sharing Platforms.
- **Video sharing platforms (VSPs):** they provide an ideal platform through which to promote audio-visual content. Search functions within these sites mean that content can easily be found, and file size limits are typically larger than on most other online platforms.
- **File hosting platforms:** File hosting or pasting sites are used to store content such as videos, images, and audio files. They are also used to aggregate information, such as lists of URLs to further content stored elsewhere.
- **Gaming-related platforms:** gaming platforms can be used to radicalise and recruit, and to propagate ideologies. Some gaming platforms also have chat functions that can be used to communicate, plan attacks and events, as well as to stream attacks.
- **Terrorist (and violent extremist) operated websites:** Websites that are run by terrorist and violent extremist groups or their supporters with the intended purpose of serving a terrorist or violent extremist group or network's interests. These play an important role in the online terrorist and violent extremism ecosystem, often acting as a centralised source of content that may have been removed from social media platforms or messaging platforms. Unlike most content on messaging apps or social media sites, content found on these sites is often indexed by search engines. Unlike accounts on third-party platforms like Facebook, Twitter, or Telegram, terrorists and violent extremists can control content on websites, as individual posts or pieces of content are not liable to content moderation.

Source: Tech Against Terrorism, "GIFCT Technical Approaches Working Group – Gap Analysis and Recommendations for Deploying Technical Solutions to Tackle the Terrorist Use of the Internet", July 2021, available at https://gifct.org/wp-content/uploads/2021/07/GIFCT-TAWG-2021.pdf

Online services' response to TVEC and government demands to do more

Most online content-sharing services prohibit in their Terms of Service or Community Guidelines the use of their technologies to support or engage in terrorist and violent extremist activities, and take preventive and reactive measures when violations of their rules occur, including warnings, content removal, account suspensions and permanent bans. However, the role that certain platforms have played in the dissemination of TVEC has led to demands to do more. In a 2018 speech, former United Kingdom Prime Minister Theresa May urged online content-sharing services "to move further and faster in reducing the time it takes to remove terrorist content online", noting that such content should be "removed automatically"

(May, 2018[7]). In a similar vein, in its 2018 Recommendation on measures to effectively tackle illegal content online, the European Commission listed a number of measures to eradicate the uploading and sharing of TVEC online, including the use of automated means (European Commission, 2018[8]). Calls to increase efforts to limit the spread of TVEC online have also been voiced in international fora (G20, 2019[9]; G7, 2019[10]; G20, 2017[11]; Christchurch Call, 2019[12]). As explored in Section 4 of this report, many OECD countries have introduced bills and otherwise implemented initiatives to help address the problem of TVEC online.

Some online content sharing services were already using automated techniques and coordinating actions to curb the dissemination of TVEC before the aforementioned calls and recommendations. Given the scale of content they must moderate and the array of conduct that may violate their Terms of Service – from child sexual abuse and hate speech to bullying and spam – services like Facebook, Twitter, and YouTube have been pioneers in the use of algorithmic techniques for identifying and removing content in violation of their rules. Also, as discussed in Section 3 of the first and second benchmarking reports, Facebook, Microsoft, Twitter, and YouTube formed the Global Internet Forum to Counter Terrorism (GIFCT) in 2017 with an aim to prevent terrorists (and now violent extremists, as well) from exploiting digital platforms (GIFCT, n.d.[13]). Under one of the GIFCT's initiatives – the Hash Sharing Consortium – its members can avail themselves of a database of "hashes" or "digital fingerprints" of known TVEC that has been already detected and removed by at least one participating company. Empowered by this database and in accordance with their internal policies, consortium members can swiftly prevent the same TVEC from being re-uploaded, at the same time making unavailable any copies of said TVEC that exist on other participating services (OECD, 2020[14]) (OECD, 2021[15]). In the context of the COVID-19 pandemic, which resulted in a shortage of human moderators due to lockdowns and work-from-home measures, many platforms increased their reliance on automated monitoring systems to flag and remove problematic content, including TVEC (Lymn, 2021[16]).

Those initiatives seek to achieve a common goal: to remove TVEC as fast as practicable, and to the extent feasible, to prevent it from being viewed in the first place. However, international calls for action such as those of the G20, the G7, and the Christchurch Call, have also highlighted the need for online service providers to be more transparent and accountable with regard to TVEC. The concern is not only whether the companies are making adequate progress in countering TVEC. It is also that, "by simply prohibiting more speech than the law does", platforms may be using their Community Guidelines to simplify their moderation efforts and avoid conflicts with governments, whilst governments may avoid legal challenges if removal of lawful content is attributed to private – instead of state – action (Keller and Leerssen, 2020[17]). Also, it is feared that overly intrusive monitoring of content to remove TVEC may impinge upon individuals' fundamental rights and freedoms such as the right to privacy, data protection and freedom of expression (Article 19, 2020[18]), and it is questioned whether the processes leading to decisions to block or remove content respect due process guarantees and wider rule of law values. Some observe that companies' appeal processes have been highly uneven, with individuals waiting for months to have their complaints resolved and sometimes never getting any response (Guillemin, 2020[19]). Transparency reporting has the potential to shed valuable light on all of those concerns.

However, online platforms historically have had little incentive to disclose information on their content moderation practices; compiling data on detection and actioning can be time-consuming and cumbersome, and any disclosure – especially one involving the admission of error – could be used against them in court or in the press (Keller and Leerssen, 2020[17]). Nonetheless, some online content-sharing services firms have stood out for their commitment to provide greater transparency in setting their Terms of Service and Community Guidelines, as well as in the manner in which they enforce them. Facebook, YouTube, Twitter and Automattic, for example, were amongst the first ever to issue transparency reports on TVEC, and as can be seen in Section 2 of this report, the number of companies doing so is growing.

Increased transparency by a growing number of content sharing services is likely to benefit individuals, the reporting companies and society at large. The appearance of TVEC reduces user trust in the Internet and harms the business models and reputations of the affected companies (European Commission, 2018[2]). By providing clear information on how they tackle this issue, conversely, content-sharing services can demonstrate their commitments to protect their users against exposure to TVEC, ensure an overall safer online environment, and safeguard human rights including freedom of expression, thus taking an important step to restore the trust and safety that TVEC has undermined. Moreover, without information on companies' actual moderation practices and capabilities, debates on appropriate regulatory or legislative measures become inherently speculative. As a result, ensuing laws and regulations can be burdensome for companies and their users, and also fail to achieve the lawmakers' intended goals (Keller and Leerssen, 2020[17]). Reliable, consistent and comparable information is essential to engage in evidence-based analysis and sound policy- and rule-making.

The evolution of the OECD's annual benchmarking report on TVEC

Against this background, the OECD launched a multi-faceted project to develop a framework and standardised set of metrics for voluntary transparency reporting on TVEC. The project included the elaboration of annual benchmarking reports that provide additional context and motivation for the transparency framework. The first report, Current Approaches to Terrorist and Violent Extremist Content Among the Global Top-50 Online Content-Sharing Services, published in August 2020, provided a snapshot of the TVEC-related policies and procedures of the world's top-50 most popular (i.e. those with the largest user bases) online platforms and other online content-sharing services (the "Services"). It identified commonalities, developments and trends in their approaches. An important aspect of this inquiry was whether the Services issued transparency reports on TVEC, the types of metrics they reported if so, their calculation methodology, their degree of comprehensiveness, the frequency with which they were issued, and other relevant factors.

The second report, Transparency Reporting on Terrorist and Violent Extremist Content Online, and Update on the Global Top-50 Content-Sharing Services, published in July 2021, focused on the degree to which the Services' approaches to tackling TVEC online had changed and evolved over the course of one year. Special emphasis was placed on whether there was more or less clarity in how the Services defined TVEC and the enforcement procedures they follow to address it, whether the number of Services that publish transparency reports on TVEC had changed, and what metrics those reports included.

This third edition is broader in scope, as it takes into account the "displacement effect" resulting from more aggressive content moderation by a number of large online platforms. An illustration of the displacement effect is provided by a 2014 campaign by Twitter to eradicate IS supporters from its service. It was estimated that prior to that campaign there were at least 46,000 and potentially over 90,000 IS supporter accounts on Twitter, with an average base of followers of 1,004 (Berger and Morgan, 2015[20]). However, after the crackdown, IS supporters on Twitter were forced to dramatically increase their efforts to rebuild the IS network, thus neglecting the dissemination of propaganda, recruitment and other activities (Berger and Morgan, 2015[20]). Thereafter, IS Twitter activity was "reduced to tactical use of throwaway accounts for distributing links to pro-IS content on other platforms, rather than as a space for public IS support and influencing activity" (Conway et al., 2019[21]). A tactical use of Twitter is out-linking, which in this case means posting tweets with non-Twitter URLs that take the user to pro-IS content hosted on other platforms and sites. Research has shown that around half of these hosting sites are small- or micro-platforms (Tech Against Terrorism, 2019[22]). Also, as IS supporters lost their foothold on Twitter, they flocked en masse to the encrypted messaging platform Telegram, turning it into their formal communication channel to share "official" IS content (Prucha, 2016[23]). Other encrypted services have also been used to communicate and plan terrorist attacks (BBC, 2017[24]).

Similarly, the removal of far-right organisations and activists from Facebook, Twitter, YouTube, Cloudflare, and other platforms following incidents like the 2017 Unite the Right rally in Charlottesville, Virginia (United States) has resulted in the emergence and growing consolidation of an ecosystem of alternative platforms – the so-called Alt-Tech (Jasser and McSwiney, 2021[25]). Alt-tech platforms like Gab have experienced substantial growth, especially after the January 2021 attack on the United States Capitol and the subsequent deplatforming of Parler which saw Gab gain an estimated 10,000+ new users per hour for a time (Brandt and Dean, 2021[26]). The Alt-Tech's highly lax or completely absent content moderation practices provide a safe space for people sharing extremist views and allow the proliferation of hateful and extremist material, speech and conspiracies, thus attracting users banned from mainstream platforms (Jasser and McSwiney, 2021[25]) (Scrivens et al., 2021[27]).

Therefore, stronger responses by large platforms to address TVEC have resulted in switching patterns on the part of terrorist and violent extremist groups and individuals. These bad actors tend to find shelter in small- or micro-platforms that lack either the financial and technical resources to make their services unappealing to them, or the will to do so. These platforms are either security-oriented, encryption-based platforms that afford high degrees of privacy, anonymity and confidentiality, or fringe, Alt-Tech platforms that seek to capitalise on deplatforming efforts by mainstream platforms to expand their user base and provide infrastructure to free speech fundamentalists, provocateurs, white supremacists and other extremist groups (Donovan, Lewis and Friedberg, 2019[28]).

The abovementioned developments suggest that an effective response to eradicate TVEC online requires broadening the focus from just the largest, most popular platforms to the entirety of the platform "ecology" (Conway et al., 2019[21]). Only by doing so can patterns of interaction and TVEC distribution amongst the ecology's components be identified, and counter-TVEC strategies aimed at disrupting their ties of complementarity be devised accordingly. Therefore, in addition to providing an update on the global top-50 Services' approaches to countering TVEC online, this edition in the benchmarking series explores the TVEC-related policies and procedures of the world's top-50 online platforms and other content-sharing services that terrorist and violent extremist organisations and individuals rely upon the most (the "TVEC-intensive Services").

Like the first two editions, this report provides an objective and factual snapshot of the Services' – and now also of the TVEC-intensive Services' - current policies and procedures for combatting TVEC. No opinions on the merits of said policies and procedures are expressed, nor are any recommendations about them made. This report is intended only to provide an evidence base for understanding the Services' and the TVEC-intensive Services' approaches (if any) to thwarting TVEC and determining the extent to which their implementation is transparent and accountable. Importantly, this report also supports efforts led by the OECD, in consultation with a group of expert stakeholders from member countries, online content-sharing services, civil society and academia, to develop a consensus-driven framework and set of metrics for voluntary transparency reporting on TVEC by content-sharing services (see the Voluntary Transparency Reporting Framework pilot portal).

The remainder of this report is structured as follows: Section 1 explains the report's research methodology and scope, detailing how it relates to the two previous benchmarking reports, as well as the limitations of the TVEC-intensive Services' selection criteria, Section 2 summarises the first two benchmarking reports' main findings and explores the development and evolution of the Services' approaches to tackling TVEC online over the past year, Section 3 explores the TVEC-intensive Services' TVEC-related policies and procedures, identifying variances, similarities and trends, and Section 4 concludes with an overview of the main legal and regulatory instruments and proposals concerning TVEC in OECD jurisdictions, detailing the manner in which they have progressed since the second benchmarking report.

1. Scope, Methodology and Research Design

The main objective of the first benchmarking report was to determine the state of play amongst the world's top-50 Services with regard to their policies, procedures and practices relevant to TVEC. The Services included social media platforms, online communications services, file sharing platforms, and other online Services that enable the uploading, posting, sharing and/or transfer of digital content and/or facilitate voice, video, messaging or other types of online communications. The Services were chosen on the basis of the size of their user bases - i.e. the extent to which they are "popular" - as it was assumed that TVEC showing up on them was bound to have a greater audience reach, which is a key element terrorist groups focus on when choosing their platforms (Tech Against Terrorism, 2019[22]). One year later, the second benchmarking report followed the same approach and tracked the evolution of the Services' approaches to tackling TVEC online over that one-year period.

Although the extent to which a Service is popular determines the size of the audience any TVEC disseminated on it may reach, popular Services are not necessarily those that are most exploited by terrorist and violent extremist groups and individuals. As explained in the Introduction, TVEC eradication campaigns by large online platforms have caused a displacement effect whereby terrorist and violent extremist groups' supporters switch to smaller, security-oriented or fringe platforms to continue with their TVEC dissemination and violent mobilisation efforts. This report acknowledges this reality, as it addresses the TVEC online problem as a whole instead of focusing only on the most popular platforms.

Thus, this report is concerned with the approaches to TVEC of services included in two distinct rankings: the top-50 most popular services (the "Services"), and the top-50 services that are exploited the most in furtherance of terrorist and violent extremist aims (the "TVEC-intensive Services"). In particular, the developments and trends in the TVEC-related policies and procedures of the former group since the publication of the second benchmarking report are identified in Section 2. Special focus is placed on whether there is more or less clarity in how the Services define TVEC and associated concepts such as terrorist organisations or hate speech, the extent to which the enforcement procedures they follow to detect and address such content is clear and transparent, and whether the number of Services that publish transparency reports on TVEC has changed. When a Service starts or continues publishing transparency reports on TVEC, particular attention is paid to the metrics those reports include, as well as to their calculation methodologies.

As in the first two benchmarking reports, given the absence of a common metric that could establish the popularity of all the surveyed Services, a two-step approach to rank them was followed. First, the Services were organised into three categories:

1. social media, video streaming services and online communications services;
2. cloud-based file sharing services; and

3. an "other" category, which includes a content management service and an online encyclopaedia.

Then, within each category, the most popular Services were chosen. To determine popularity, the following metrics were employed:

- Social media platforms, video streaming services and online communications services were chosen 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[2], 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 chosen 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. There are some noteworthy changes in this list as compared to that included in the second benchmarking report. The coronavirus pandemic and associated lockdowns forced millions of people around the world to work from home, thus triggering the explosive growth of video conferencing and collaboration apps like Zoom and Microsoft Teams, which made it into the list and displaced 4chan and Meetup as a result.

In turn, just as in the case of popularity measurement, finding one common metric to determine which platforms and services terrorist and violent extremist groups use the most proved impossible. A strictly quantitative approach – i.e. counting the pieces of TVEC that appear on the services to be ranked within a certain period of time - was originally contemplated. However, there are two main problems with that approach. First, reliable, consistent and accurate data on volumes of surfaced TVEC on different platforms and websites is painfully difficult to collect, even when assisted by sophisticated website analytics and scraping tools. Some platforms use end-to-end encryption protocols, so not even the platform operator knows that volume. Moreover, most platforms have privacy settings which prevent third-party scanning of user profiles that are not set as public; hence, content in those profiles is not accessible, and by extension any TVEC there cannot be detected. Also, terrorist and violent extremist groups employ multiple content moderation circumvention techniques - e.g. mirroring, language amendments, content editing and repurposing (for example, posting content resembling legitimate news or use of hashtags like #savethechildren to conceal TVEC) and outlinking (i.e. posting URLs featuring TVEC hosted elsewhere (Tech Against Terrorism, 2021[29]). Therefore, a certain volume of TVEC is bound to go uncounted. At any rate, as we have seen in the past two benchmarking reports, many companies are not particularly forthcoming with regard to reporting on TVEC, and even when they are, they tend to follow different reporting approaches. As a result, this data cannot be easily compared and relied upon to rank TVEC-intensive Services.

Secondly, and most importantly, the volume of TVEC appearing on a given platform is not necessarily the most decisive element determining that platform's impact on and magnitude in the overall TVEC online landscape. A website may be fully dedicated to the dissemination of TVEC, but if it receives very few visits or has a very small userbase, its contribution to TVEC dissemination will likely be minimal. Also, popular platforms tend not to be used anymore to reach large audiences directly (e.g. uploading a terrorist video on YouTube), but rather to refer to other websites or platforms where TVEC is stored (outlinking). Put simply, measuring TVEC prevalence – the proportion of views of removed TVEC versus the total content viewed over a certain period of time – is as important as, or even more important than, measuring the quantity of TVEC uploaded and shared.

These two different aspects of the extent to which TVEC online is impactful cannot be captured by one common metric. Determining which platforms and services are exploited the most by terrorist and violent extremist groups and their supporters in a way that captures the magnitude of their contribution to TVEC availability and dissemination is thus a complex exercise which requires, on one hand, consideration of the different ways in and purposes for which terrorist and violent extremist groups rely on digital technologies, and on the other hand, an assessment of the likely reach, or success, of such groups' TVEC dissemination practices. Depending on the available data and the type of service used for TVEC-related purposes (e.g. mainstream social networks or cloud-based file sharing platforms), some metrics capture better the magnitude of a service's impact on TVEC availability and dissemination than others.

To address these challenges and be able to rank the TVEC-intensive Services in a manner that best depicts their actual contribution to the overall TVEC online landscape, the OECD partnered with SITE Intelligence Group (SITE), a company with over two decades of experience in providing governments and institutions across the world with verified, actionable intelligence and analysis on designated terrorist and violent extremist groups, as well as the larger movements from which these groups originate . After considering many approaches to ranking the most TVEC-intensive Services based on the data that SITE was able to gather for the January-December 2021 period (which includes only jihadi and far-right groups), the top-50 TVEC-intensive Services were divided into three categories: mainstream TVEC-intensive Services, file-sharing TVEC-intensive Services and far-right-focused TVEC-intensive Services. The platforms and websites included in the ranking were selected by SITE's expert analysts as the most prevalently shared/used ones on the spaces they monitor.

More specifically, the services in the first two categories were ranked based on the number of URLs shared on Telegram in 2021 that linked to TVEC and TVEC-related posts on those services. Telegram has historically been the primary hub of jihadi activity (Katz, 2019[30]), and recently has gained significant popularity amongst far-right groups and their supporters, as well (Katz, 2020[31]) (Quinn, 2021[32]). Thus, the URLs of different websites and platforms found on Telegram linking to TVEC-related posts, group chats, threads and TVEC *per se* (as well as mobile and condensed variants of those URLs) give a telling look into the intensity with which such websites and platforms are exploited by jihadi and far-right groups/supporters. The URL data was acquired via SITE's SourceFeed database, which contains years of archived and contextualized[3] TVEC from jihadists and far-right extremists alike, ranging from chat logs, web page HTMLs, multimedia releases, and other artifacts. More concretely, the URLs data was tallied using a specific feature of SourceFeed called SearchFeed, with which SITE analysts were able to search each platform's domain. These messages were filtered specifically to Telegram, leaving a pool of nearly 90 million saved messages to search from 2021 alone.

The number of URLs metric takes into account both the lack of reliable and consistent data on the volume of TVEC surfaced on the ranked services and an important way in which such services are used to disseminate TVEC: outlinking. Terrorist and violent extremist posts, group chats, threads and discussions found on mainstream social media and messaging platforms often contain URLs to content hosted elsewhere; this content ranges from unambiguous TVEC to non-TVEC that indirectly contributes to artifacts and conversations that amount to TVEC. In turn, file-sharing services – especially small or fringe services – are sometimes used for TVEC storing and backup. Thus:

- The first category is composed of the mainstream platforms that are most exploited by jihadi and far-right groups, as defined by the number of URLs (as well as mobile and condensed variants of those URLs) found on Telegram during 2021 linking to TVEC-related posts, group chats, threads and discussions available on these platforms.
- The second category is comprised of such platforms that are used by jihadi and far-right groups for the purposes of uploading, storing and sharing files containing TVEC. Like the mainstream category,

these services were ranked based on the number of URLs (as well as mobile and condensed variants of those URLs) linking to TVEC stored on them found on Telegram during 2021.

Lastly, the third category includes the top platforms and websites that are either created by, predominantly exploited by, or accommodating to far-right extremists. This category does not include jihadi sites, as jihadi groups are largely unable to maintain websites because online service providers remove them far more aggressively than they do far-right sites, and far-right groups tend not to be included in terrorist lists. Thus, for example, Parler's Legal Guidelines provide that "terrorist organizations officially recognized as such by the United States are forbidden from using Parler, as is anyone—including state actors—recruiting for them"[4] – yet Parler is notorious for providing a safe haven for extremist groups and supporters and the site remains up and running after being "deplatformed" by Amazon. Some of the platforms and websites included in this category strongly prioritize freedom of speech and tend not to moderate content. Others are specifically designed to spread violent extremist views and cater to TVEC. Therefore, a metric that captures their usage, or popularity – and by extension the prevalence of TVEC on them – better depicts their significance in the overall TVEC landscape than the number of URLs metric (which provides no information on the number of times the TVEC to which the URLs refer is seen). Accordingly, the platforms and websites in this category are ranked on the basis of the number of visits and unique visitors they had during 2021. The data was collected via a web analytics tool called Semrush.

It is important to note that jihadi and far-right groups do not comprise, nor perfectly overlap with, the totality of terrorist and violent extremist groups and individuals. The focus on these two groups reflects the available data. Similarly, whilst the Telegram-centred approach in particular and the aforementioned research methods in general help to account for the different ways in which platforms and websites fit into the overall (jihadist and far-right) TVEC landscape, they are an imperfect proxy for reality. They were chosen because reliable and consistent data on a per service basis is not yet available. Moreover, the following methodological limitations and nuances must be considered:

- Duplicate postings: In a small percentage of cases, SITE's SourceFeed database may yield multiple versions of the same pages. While these instances fall well within the contemplated margin of error and do not affect the overall platform ranking, it is important to note this possibility. (SourceFeed sometimes saves the same pages multiple times to keep a record of media and discussions that might be deleted by the monitored terrorist and violent extremist communities).
- Different contexts of ICT platform usage**:** The nature of these shared URLs varies from overtly terrorist and violent extremist propaganda to a mainstream news segment on YouTube that relates to the subject of a given terrorist and violent extremist discussion. However, even when the content associated with reshared URLs is not explicitly TVEC (e.g. a mainstream news segment about a topic of interest), each instance nonetheless reflects a critical way in which terrorists and violent extremists embrace mainstream platforms — and how non-TVEC content on those platforms is used toward terrorists and violent extremists' creations of larger TVEC artifacts and discussions. SITE's analysts were thus required to carry out a qualitative analysis of this content to determine its contribution to TVEC availability and the significance in the TVEC landscape of the service hosting it.

Summing up, the data relied upon to rank the TVEC-intensive Services emphasises these Services' significance in the TVEC landscape and relative popularity versus others; it does not, however, reveal the number of times that these Services are used to access and disseminate TVEC.

A list of the world's top-50 TVEC-intensive Services is included in Annex C. In turn, the TVEC-relevant policies, procedures and practices of the TVEC-intensive Services are presented in Section 3, highlighting commonalities, variances and trends in them.

The review of the Services' and the TVEC-intensive Services' approaches to combatting TVEC online was completed following three main steps.

a) First, the standardised profile template produced in the first benchmarking report[5] was used to profile each Service and TVEC-intensive Service. One profile per entity was produced based on its publicly available terms of service (ToS), community guidelines and policies, blogs, service agreements and other official information ("governing documents")[6]. The Services were contacted and asked to provide feedback on the accuracy of their profiles and any additional relevant information. The same step will be taken in respect of the TVEC-intensive Services.

b) Second, the Services' profiles were updated based on the Services' responses. These updated profiles are included in Annex B. The same process will be followed for TVEC-intensive Services upon receiving their feedback. The interim versions of the TVEC-intensive Services' profiles are included in Annex D.

c) Third, the developments and changes in the Services' approaches to TVEC over the course of the last year, as well as the current state of play of the TVEC-intensive Services' procedures and efforts to combat TVEC online, were identified and summarised.

Key aspects in the Services' and TVEC-intensive Services' approaches to TVEC that are surveyed in this report include:

- definitions of terms like terrorist/terrorism and violent extremist/violent extremism;
- 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;
- consequences for user breaches of terms of service/community guidelines and standards; and
- voluntary issuance of transparency reports (TRs) concerning TVEC, including their content, methodology and frequency.

2. Updated Commonalities, Developments and Trends in the Top-50 Services' Approaches to TVEC

Different descriptions of TVEC and related concepts remain, but more definitional clarity is emerging

The first benchmarking report found dissimilar approaches in the Services' definitions of TVEC and related concepts, with few Services providing definitions that allow a clear accurate understanding of what type of content is prohibited, as well as what is considered a terrorist/extremist group or organisation. The second benchmarking report found that no significant changes had taken place over the course of one year.

This time, conversely, whilst different descriptions of TVEC and related concepts remain, a trend towards more definitional clarity is emerging. Not only do more Services provide comprehensive definitions and examples of TVEC, but fewer Services consider posting TVEC, hateful speech, hateful content and/or violent or graphic content to be the same category of policy violation.

Table 1 – Services' approaches to defining TVEC and related concepts

Approach	1st benchmarking report	2nd benchmarking report	3rd 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[7]	6[8]	11[9]
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[10]	21[11]	19[12]
Services that include TVEC within the same category as hate speech, hateful content and/or violent or graphic content	15[13]	13[14]	5[15]
Services that use broad and/or vague descriptions of prohibited conduct, which descriptions can be interpreted as supersets encompassing TVEC	16[16]	15[17]	15[18]

Services that were already providing detailed definitions of TVEC and related concepts made an effort to improve their definitions and make their approaches to TVEC more transparent and comprehensible. For example, Facebook overhauled its policy on Dangerous Individuals and Organisations (which also applies to Instagram and Messenger), creating a three-tier system with different levels of content enforcement, with Tier 1 being the most aggressive on account of the risk it poses. Tier 1 focuses on entities that engage in serious offline harms, including organising or advocating for violence against civilians, repeatedly dehumanising or advocating for harm against people based on protected characteristics, or engaging in systematic criminal operations (e.g. terrorist, hate and criminal organisations). Tier 2 focuses on entities that engage in violence against state or military actors ("violent non-state actors"), whilst Tier 3 focuses on entities that may demonstrate strong intent to engage in offline violence in the near future, but have not necessarily engaged in violence to date or advocated for violence against others based on their protected characteristics (e.g. militarised social movements, violence-inducing conspiracy networks, and individuals and groups banned for promoting hatred) (Facebook, n.d.[33]). The explanation of each tier is highly detailed, with clear definitions of each type of entity and multiple examples of prohibited content and conduct falling within each tier. Facebook has clarified that the expansion of its Dangerous Individuals and Organisations policy was motivated by the need to address violent conspiracy networks such as QAnon (Facebook, 2020[34]).

In a similar vein, Discord now provides a concise yet very clear definition of violent extremism and explains its approach to this conduct and TVEC more broadly in simple and understandable terms. Noting that violent extremism is nuanced and the ideologies and tactics behind them evolve fast, Discord explains that it does not try to apply its own labels or identify a certain type of extremism. Instead, Discord evaluates user accounts, servers, and content based on common characteristics and patterns of behaviour, such as

embracing radical and dangerous ideas that are intended to cause or lead to real-world violence; the targeting of other groups or individuals who are perceived as enemies, usually based on a sensitive attribute; dismissing alternative opinions or ideas; and efforts to convince others to join a radical ideology (Discord, 2021[35]).

Another trend that can be identified is a general effort to capture content and conduct that fall outside narrow definitions of TVEC, yet has the capacity to incite attacks on specific groups, ultimately leading to real-life terrorist and violent extremist attacks. In doing so, some Services are taking steps to address what some perceive as bias in counter-TVEC efforts, which have "a disproportionate focus on Islamist extremist content" (BSR, 2021[36]). As the experience of the GIFCT shows, content moderation efforts that are guided by the United Nations Security Council's Consolidated Sanctions List are likely to focus almost exclusively on content related to al-Qaeda, the Taliban, IS or other groups designated as terrorists by the United Nations (Rasmussen and Lowin, 2021[37]). Thus, expressions of support for extremist ideologies which target minority groups – such as the white supremacist manifestos released by the perpetrators of the Christchurch and El Paso attacks (Anti-defamation League, 2021[38]) – can escape content moderation if not specifically addressed otherwise. To address this blind spot, Twitch expanded its Hateful conduct policy, listing highly detailed categories of content and behaviour that falls within it, and several examples for each category. One category is "content that encourages or supports the political or economic dominance of any race, ethnicity, or religious group, including support for white supremacist/nationalist ideologies" (Twitch, 2021[39]), a category which can also be found in Pinterest's Hateful Activities policy (support for white supremacy, limiting women's rights and other discriminatory ideas). Twitter also expanded its Hateful Conduct policy with a view to address the risks of offline harm, explaining that "research shows that dehumanizing language increases that risk" (Twitter, 2020[40]). Relatedly, Tumblr has observed that its recently introduced Election Integrity Policy in Tumblr's Community Guidelines has helped address some far-right violence extremism on Tumblr[19].

Similarly, TikTok elaborated further on its prohibition of Hateful Behaviour, now defining "hateful ideologies" as "those that demonstrate clear hostility toward people because of their protected attributes" (which are listed). This update clarifies that hateful behaviour and violent extremism may overlap: if they actually attack people based on protected characteristics, individuals or organisations which embrace a hateful ideology – and therefore engage in hateful behaviour – also fall within TikTok's prohibition on organised hate, which is part of TikTok's policy on Violent Extremism (TikTok, 2021[41]). VK also clarified its approach to "Hostile or Hate Speech", noting that its moderators look for certain signs in the content to apply this prohibition, such as animosity based on certain characteristics or differences; offensive behaviour, contempt toward other people's values or views; and expression of personal superiority, accompanied by a baseless and unfair attitude toward a specific individual or group of people (VK, 2021[42]). Notably, these signs are similar to those Discord relies upon to identify what it deems violent extremist content. This similarity underscores the extent to which hateful speech can overlap with violent extremist content.

Relatedly, Facebook also expanded its Violence and Incitement policy, which now includes an explicit prohibition of "any content containing statements of intent, calls for action, conditional or aspirational statements, or advocating for violence due to voting, voter registration or the administration or outcome of an election" (Facebook, 2021[43]). This type of conduct was common amongst far-right and conspiracy groups and networks like the Proud Boys and QAnon (groups banned from Facebook and Instagram under their Dangerous Individuals and Organisations policy) in the months preceding the insurrection at the US Capitol. This fact illustrates how difficult it is to draw a clear line between incitement of violence and violent extremism, and confirms that complementary policies and prohibitions must be implemented to prevent extremist groups from using online platforms to organise hateful activities that are likely to culminate in violent incidents.

The past two benchmarking reports found that, when defining and identifying a terrorist/violent extremist organisation, the Services had different approaches. After one year, no significant developments can be

noted in this regard. Broadly speaking, three markedly different approaches can be distinguished. Firstly, a few Services including Facebook, Instagram and Twitter provide their own definitions of terrorist/violent extremist organisations, breaking down this concept into different subsets such as hate organisations, violent non-state actors or 'other' violent organisations[20]. Secondly, other Services rely on government or United Nations lists of terrorist organisations including Microsoft's Services (LinkedIn, Teams, Skype and OneDrive), YouTube, Wordpress.com, Vimeo, Quora, Reddit and VK[21]. Lastly, the majority of the Services are still silent in this regard[22].

The number of services issuing transparency reports expressly addressing TVEC is increasing

The number of Services issuing transparency reports with specific information on TVEC is steadily increasing. In the first benchmarking report, this number was 5. Two years later, this number is 15.

Table 2 – Services that issue transparency reports with information on TVEC

1st benchmarking report	2nd benchmarking report	3rd benchmarking report
Facebook	Facebook	Facebook
YouTube	YouTube	YouTube
Instagram	Instagram	Instagram
Twitter	Twitter	Twitter
Wordpress.com	Wordpress.com	Wordpress.com
	Skype	Skype
	OneDrive	OneDrive
	Twitch	Twitch
	TikTok	TikTok
	Reddit	Reddit
	Discord	Discord
		Zoom
		Snap
		Pinterest
		Teams

The number of Services issuing transparency reports on TVEC could be higher. Ask.fm participates in the GIFCT's Hash Sharing Consortium and recently published a transparency report covering 2020 (Ask.fm, 2021[44]). The report features a whole section on content moderation, which discloses actions and metrics relating to violations of Ask.fm's Community Guidelines, including its prohibition of terrorist organisations and content. However, the violations of its TVEC prohibition are included with several other prohibitions in a category called "Other", so it is not possible to discern TVEC-specific numbers or statistics. The only TVEC-specific information included in the transparency report is the number of cases concerning violent extremism that Ask.fm reported to law enforcement agencies during the reporting period. Given the narrow

scope of this information and the fact that a number of specific metrics and numbers are reported for various categories of policy violations other than TVEC, Ask.fm has been left outside the group of Services that issue transparency reports on TVEC.

The fact that the number of Services issuing transparency reports with specific information on TVEC has at least tripled after two years is noteworthy, not least on account of the difficulties in assembling an initial transparency report, especially for small or new companies. Producing a transparency report involves a huge internal effort to collate data that must be replicable and accurate, typically requiring engineering, data support, and policy and legal teams for reviewing. Indeed, initial transparency reports tend not to have a granular focus on TVEC, as this may not be a particularly recurrent issue on the service, or the company may need to prioritise transparency about government requests for user data (OECD, 2021[45]) or other categories of violating content such as child sexual abuse material (GIFCT Transparency Working Group, 2021[46]). Pinterest, Snap and Discord fall within this pattern.

The production of transparency reports also depends on the relevant company's upstream content moderation processes and systems. For example, a number of platforms questioned by Tech Against Terrorism as to why they did not issue transparency reports explained that they did not report specifically on TVEC due to the lack of a dedicated reporting category in their reporting processes and ensuing content moderation workflow (Tech Against Terrorism, 2021[29]). In essence, if there is neither a policy on TVEC in place nor content moderation of this category, there is nothing to report. Therefore, a growing number of Services disclosing information on TVEC removals means that more Services are implementing specific prohibitions of TVEC and deploying dedicated counter-TVEC efforts. This development suggests that the fight against TVEC online is gaining momentum.

With this trend towards more transparency, the Services are increasingly allowing academic researchers, regulators and other stakeholders to have a basic understanding of how they apply their policies, to what extent these policies are enforced, the volume of TVEC that is detected, and the extent to which TVEC enforcement efforts are effective. In this way, the scope of the relevant problem can be clarified, and the Services' actions can be assessed in a more informed manner.

Transparency reports on TVEC are becoming more detailed

When it comes to the content of transparency reports on TVEC, two main trends can be discerned since the first benchmarking report. First, the Services that have already issued transparency reports tend to include more granular information or new metrics in subsequent releases. Second, Services that have already issued transparency reports, but not specifically about TVEC, generally supplement subsequent releases with additional information about their efforts to tackle TVEC and/or explanations of new metrics' rationales and calculation methodologies.

For example, Facebook and Instagram continue reporting the same metrics as last year, i.e. 1) how prevalent dangerous organisations violations on Facebook/Instagram were (prevalence, which refers to how often content that violates Facebook's and Instagram's standards is actually seen relative to the total amount of times any content is seen on Facebook/Instagram); 2) how much dangerous organisations content Facebook/ Instagram took action on (content actioned); 3) of the violation content actioned for dangerous organisations, how much did Facebook/Instagram find before people reported it (proactive rate); 4) how much content actioned was appealed (appealed content); and 5) the amount of content Facebook/Instagram restored after removing it (content restored). However, all of these metrics are now broken down into terrorist content and organised hate. Moreover, the proactive rate metric now also includes the percentage of actioned content reported by users. This information sheds light on the extent to which Facebook and Instagram rely on their automated tools and user reports to moderate content, and clarify how central their automated tools are in these efforts. Lastly, the restored content metric is now broken down into content restored without appeal and content restored after appeal. This granularity

elucidates the number of occasions where Facebook/Instagram proactively corrected content moderation decisions, as well as the extent to which appeal processes are successful.

Facebook and Instagram also made an effort to provide more transparency as to how their content moderation process works. In particular, they explained how their technology detects violations, how their algorithms work and are trained, how technology helps prioritise review, and how their review teams work and are trained. Further, they provided information about their process for taking down violating content, how strikes are counted, and how accounts are restricted and disabled (Facebook, 2021[47]).

Importantly, Facebook and Instagram also clarified the rationale of their prevalence metric. They observed that a piece of violating content can be published once but seen 1,000 times, one million times or not at all. Thus, measuring views of violating content rather than the amount of violating content published better reflects the impact of content violations on their platforms. A small prevalence number can still correspond to a large impact, due to the large number of overall views of content on Facebook and Instagram (Facebook, 2021[48]).

In the same vein, Twitch clarified the rationale of its content moderation enforcement. Twitch is a live-streaming service, so the vast majority of content that appears on Twitch is gone the moment it is created and seen. Therefore, its approach to content moderation must differ from that of other services that are primarily based on pre-recorded and uploaded content. 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, and content moderators then issue "enforcements" (typically a warning or timed channel suspension) for verified violations (Twitch, 2021[49]). Accordingly, "enforcements" are the main focus of Twitch's transparency report. In particular, it discloses the number of user reports for all types of violations during the reporting period (H1 and H2 2020), the total number of enforcement actions, the number of enforcement actions for reports of different categories of policy violations, including terrorism, terrorist propaganda and recruitment violations, and the number of enforcement actions for reports of these violations per thousand hours watched.

Discord, which started reporting on TVEC only last year and disclosed very limited information, expanded the scope of its transparency report dramatically. Discord now discloses the overall number of reports received, as well as the number and percentage that fell within each prohibited category (one of which is extremist or violent content); the number and percentage of the reports on which Discord took action, and the report action rate, by prohibited category; the total number of account deletions (excluding spam), by prohibited categories; the total number of server deletions, by prohibited categories; the number of accounts warned, accounts deleted after warning, servers warned, and servers deleted after warning, by prohibited categories; the number of server deletions proactively deleted and reactively deleted, by prohibited categories; and the percentage of accounts reinstated on appeal, by prohibited categories. Moreover, Discord included in its last transparency report real-life examples of its ban appeal processes, including an example of an appeal of a violent extremism violation. These examples clarify the manner in which appeals are assessed, and what the relevant criteria for reaching a given decision are.

Similarly, both TikTok and Reddit expanded the scope of information on TVEC they provide as compared to their first transparency reports (which were issued last year). TikTok now reports the percentage of videos removed by removal reason (including violent extremism, hateful behaviour and violent and graphic content); the proactive removal rate, the removal rate before any views and removal within the 24 hours rate by removal reason (including violent extremism, hateful behaviour and violent and graphic content). Proactive removal means identifying and removing content that violates the terms and conditions (also known as "violative" content) before it is reported to TikTok. Removal within 24 hours means removing the video within 24 hours of it being posted. In turn, Reddit now reports the number and percentage of Content Policy violations removed by subreddit moderators and by Reddit administrators divided by categories of violations (including Hateful Content and Violent content, the latter category including US-designated

foreign terrorist organisation content), also divided by whether the violations were surfaced via user reports or automation; the number of subreddits removed by categories of violations; the number and percentage of posts removed by categories of violation, the comments removed by categories of violation, the private messages removed by categories of violation; and the number of accounts sanctioned by categories of violations.

Twitter is also now reporting the proactive rate of accounts actioned for violations of its policy on terrorism and violent extremism. In addition, Twitter introduced a new metric called "impressions", which captures the number of views a Tweet received prior to removal. An impression is defined as any time at least half of the area of a given Tweet is visible to a user for at least half a second (including while scrolling). This new metric, however, is not broken down into categories of violations, so it is not possible to identify specific impressions prior to TVEC removals.

In a similar vein, YouTube continues reporting the same metrics as last year, but now it provides additional insights on its "violative view rate" (VVR), i.e. an estimate of the proportion of video views that violate YouTube's community guidelines in a given quarter (excluding spam). YouTube observes that to calculate this metric, it first takes a sample of all videos that have been viewed on YouTube. The videos in that sample are then sent for review, and its teams determine whether each video does or does not violate its Community Guidelines. YouTube then use the aggregate results to estimate the proportion of views on YouTube that violate its Community Guidelines. The VVR metric is reported with a 95% confidence interval. This means that if measurement were performed many times for the same period of time, YouTube would expect the true metric to lie within the interval 95% of the time. However, just like Twitter's impression metric, YouTube's VVR is not broken down into categories of violations, so it is not possible to discern the proportion of content in violation of YouTube's prohibition on terrorist/violent extremist content that was viewed before removal.

Similarly, Snap's last transparency report also includes a VVR, i.e. the proportion of all Snaps (or views) that contained content that violated Snap's Community Guidelines during the reporting period, following Twitter's and YouTube's approach (the rate is not broken down into categories of violations). In addition, Snap's last transparency report included TVEC-specific information for the first time, in particular, the number of content reports, content enforced, turnaround time and unique accounts enforced broken down by type of policy violation (which include threatening / violence / harm and hate speech); and the number of account removals for violations of Snap's prohibition of terrorism, hate speech and extremist content.

Relatedly, in its first transparency report including data on content moderation for policy violations, Pinterest reports something akin to a VVR, i.e. the reach of policy-violating Pins (content with a message), which is represented as the percentage of Pins seen by 0 people, <10 people, 10-100 people and 100+ people. Unlike the approach followed by Twitter, YouTube and Snap, however, the reach of violating pins is broken down into types of policy violations, including hateful activities and violent actors. Pinterest's last transparency report also includes the number of actioned used reports, by type of policy violation; the number of distinct images and Pins deactivations, by type of policy violation, the number of board deactivations, by type of policy violation; the number of account deactivations, by type of policy violation; and the number of account appeals and reinstatements, by type of policy violation. Likewise, Zoom also published its first transparency report with TVEC-specific data, disclosing the number and percentage of actioned reports by issue type (which includes terrorist or violent extremist groups).

Last but not least, Automattic (Wordpress.com's parent company) and Microsoft continue reporting the same TVEC-specific metrics featured in their last year transparency reports, although it is worth mentioning that Microsoft's report now also encompasses Teams, in addition to Skype and OneDrive.

Some degree of convergence in transparency reports on TVEC is emerging

Whilst variance amongst the Services' reporting approaches remains, there has been a degree of convergence with regard to six content moderation aspects: proactive detection, the actions that follow a finding that there has been a TVEC violation, the counting of TVEC violation reports, TVEC-related appeals, reinstatement of content or accounts, and views of violating content. Table 3 below depicts this convergence when at least two Services are reporting metrics on one of these content moderation aspects.

Table 3 – Convergence in 6 aspects of TVEC transparency reporting

Content Mod. Aspect (down) / Name of Service (right)	Facebook / Instagram	YouTube	Twitter	Wordpress.com	Skype / OneDrive / Teams	Twitch	Discord	Reddit	TikTok	Zoom	Pinterest	Snap
Proactive detection	Proactive Rate		Proactive Rate		Proactive Rate		Proactive Rate (servers)	Content removal via automation (Number and %)	Proactive Rate			
Actions following a TVEC violation finding	Content Actioned (Number)	Channels Removed (Number) Videos Removed (Number) Comments Removed (Number)	Accounts Actioned (Number) Content Removed (Number)	Actioned IRU Notices (Number)	TVEC Actioned (Number) Accounts Suspended (Number)	Enforcement Actions (Number)	Actioned Reports (Number and %) Account Deletions (Number) Server Deletions (Number)	Content Removed (Number and %) Subreddits Removed (Number and %) Posts Removed (Number and %) Comments Removed (Number and %) Private Messages Removed (Number and %) Accounts Actioned (Number)	Videos Removed (%)	Actioned Reports (Number and %)	Actioned Reports (Number) Board Deactivations (Number) Account Deactivations (Number) Pins Deactivations (Number)	Content Enforced (Number) Accounts Enforced (Number)

Reports counting			Accounts Reported (Number)		Accounts Reported (%)		Reports Receive (Number and %)					Content Reports (Number)
Appeals	Appealed Content (Number)							Appeals (Number)			Account Appeals (Number)	
Content or Accounts Reinstated	Restored Content (Number)				Reinstated Accounts (%)		Reinstated Accounts (%)	Appeals granted (Number)			Account Reinstatements (Number)	
Views of violating content	Prevalence	VVR*	Impressions*								Pins Reach	VVR*

*Metric reported for all policy violations, as opposed to TVEC-specific

A growing number of Services are placing particular emphasis on the impact that violating content causes when it is viewed, as opposed to when it is merely posted. As seen in the preceding section, Facebook and YouTube articulated the rationale behind their Prevalence and VVR metrics in detail, and Twitter, Pinterest and Snap introduced metrics on this aspect of transparency reporting for the first time. Currently, only Facebook and Pinterest are reporting TVEC-specific percentages in this regard; however, on account of certain regulatory developments (see Section 5 of this report), YouTube, Twitter, Snap and more Services may soon begin to break down this metric into different types of policy violations, including the posting of TVEC.

Second, on its face, convergence regarding the actions that follow a finding of TVEC violation seems to be limited, as not all the reporting Services included in the table above provide the same or similar metrics on this crucial aspect. In the context of transparency reporting, there is agreement on the fact that, if Internet-based services could be compared to one another in terms of their metrics, this would facilitate academic research, regulation and more informed debate (GIFCT Transparency Working Group, 2021[46]). However, full standardisation of reporting metrics is unlikely to be achievable, as the Services are hardly homogeneous in their functionalities, purposes, operational properties, content/use policies and how they moderate content or otherwise enforce their policies.

In fact, whilst there is significance variance as to the degree of detail and metrics that Services having largely the same or similar business model report, the metrics the Services report in particular on the actions that follow a TVEC violation finding tend to be to some extent consistent with the functionalities they provide, the manner in which they are used, and their approach to content moderation. For example, as noted above, Twitch is a live-streaming service in which the majority of the content is by definition fleeting. Thus, it makes sense that its transparency report focuses on enforcement actions of verified violations instead of, for example, content removals, as the “enforcements” metric reflects better the platform’s functionality and moderation approach. Similarly, the metrics reported by Zoom, a video conferencing platform with little to no content being posted on it, are also consistent with its functionality and content moderation system[23]. Likewise, the metrics reported by Services such as Facebook/Instagram, YouTube, Twitter, TikTok, Pinterest and Snap are in accord with the type of user-generated content they host and the manner in which their users interact on their platforms (e.g. they must hold an account or they interact in specific communities within the Service).

Thus, in order to compare and analyse granular reported metrics across reports, it may be best to place the Services into different groups or clusters based on their business models and specific features (e.g. social media platforms, live-streaming services, video-streaming services and so on).

Staff member moderators, user-moderators and automated tools: the key role of human intervention

The first benchmarking report found that the Services relied on staff member moderators, user-moderators, automated tools, or a combination of them to moderate TVEC, whilst the second benchmarking report found no relevant changes in these approaches over the course of one year. Conversely, significant variations as compared to last year can be seen in the number of Services using staff member moderators and automated tools.

Table 4 – Services' approaches to content moderation

Approach	1st benchmarking report	2nd benchmarking report	3rd benchmarking report
Services that rely on staff member moderators	40[24]	40[25]	50[26]
Services that rely on user-moderators	10[27]	10[28]	10[29]
Services that rely on automated tools	At least[30] 21[31]	At least[32] 22[33]	42[34]

The reason for the increase in the number of Services that rely on staff member moderators and automated tools is three-fold. First, academic research and media coverage has confirmed that all online services based in the People's Republic of China (hereafter 'China') use moderators and automated tools to censor content in accordance with Chinese regulatory requirements (see heading "Disclosure by Chinese Platforms" at the end of this section). Second, there is generally more transparency as to how the Services moderate content. Thus, many Services have explicitly recognised their use of moderators and automated tools. Third, there is a higher number of members of both the GIFCT and the GIFCT's Hash Sharing Consortium[35].

The second benchmarking report noted that as a result of the COVID-19 pandemic and lock-down measures, some Services like Facebook, Instagram, YouTube and Twitter faced a shortage of human moderators. As a response, they increased their reliance on automated monitoring systems to flag and remove problematic content, including TVEC (OECD, 2021[15]). This response gave rise to concerns about human rights impacts, especially on freedom of expression and due process guarantees. Moderation often requires an understanding of the context around content to determine whether or not it is harmful. However, algorithms cannot yet catch all the contextual nuances involved in borderline moderation decisions, so it is not possible to fully automate effective content moderation. Indeed, fully automated algorithmic content moderation tends to err on the side of caution, thus increasing the risk of a higher number of false positives, although the likelihood of more false negatives is also real (OECD, 2021[15]). These issues can be compounded if there are no humans to entertain appeals against automated moderation decisions.

Data included in last year's transparency reports of Facebook, Instagram, YouTube and Twitter may warrant some of the abovementioned human rights concerns. Both Facebook and YouTube roughly doubled the harmful content they removed for policy violations in Q2 2020 compared to Q1 2020 (when the pandemic started). Twitter reported a +132% increase in content removed in H2 2020 compared to H1 2020. There were also reports of activists' accounts being closed overnight without the possibility to appeal in Syria, where activists and journalists rely on social media to document potential war crimes. Meanwhile, the number of removals for violations like child sexual exploitation and self-harm on Facebook fell by at least 40% in Q2 2020 due to a lack of humans to make the necessary "tough calls" (Scott and Kayali, 2020[50]). Relatedly, anti-racism activists in France found an increase of more than 40% in unactioned hate speech on Twitter (Scott and Kayali, 2020[50]). These kinds of false positives and false negatives have clear implications for freedom of expression, the extent to which these platforms' users are exposed to undetected and/or unactioned harmful content, and due process.

Impacts on due process can also be discerned from Instagram data. In Q2 2020, around 389,000 posts potentially containing terrorist content were removed from Instagram. Whilst around 8,400 removals were appealed by users in Q1 2020, that number plummeted to zero in Q2 due to the absence of humans involved in appeals processes. In turn, during Q1 2020, 500 pieces of content were restored after a user appeal, and 200 were reinstated without an appeal. By comparison, in Q2, without an appeal system available, only 90 posts of the over 389,000 that were actioned based on the terrorist content prohibition were restored.

Overall, the developments above stress how important human intervention is to ensure fair and effective content moderation, as well as to guard users' fundamental rights and due process guarantees. The input of human moderators is still and will remain for the foreseeable required to review highly contextual, nuanced content (Cambridge Consultants, 2019[51]). Moreover, without human review there can be no appeals system, which could lead to over-censorship and chilling effects.

Notification, enforcement and appeal mechanisms and processes

The Santa Clara Principles on Transparency and Accountability in Content Moderation (Santa Clara University's High Tech Law Institute, n.d.[52]) offer a number of basic standards to ensure that content moderation does not unduly interfere with users' fundamental rights and freedoms, including notifications of enforcement decisions and the possibility to appeal them. The first two benchmarking reports found that fewer than half of the Services both notified their users in case of potential policy violations and had appeal processes in place.

Since the publication of the second benchmarking report, the number of Services that provide notifications of enforcement decisions as well as the number of Services that give users the right to appeal enforcement decisions increased to from 21 to 30 and from 24 to 31, respectively.

Table 5 – Services' approaches to notifications and appeals

Approach	1st benchmarking report	2nd benchmarking report	3rd benchmarking report
Services that have mechanisms for notifying users in case of potential violations of their ToS and other governing documents	21[36]	21[37]	30[38]
Services that have appeal processes in place in respect of content moderation decisions and other measures applied under their governing documents	23[39]	24[40]	31[41]

The remaining Services either offer no notifications/have no appeal processes implemented, or do not provide public information in this regard.

Lastly, in the first benchmarking report, obtaining a clear understanding of whether content was reviewed proactively and/or reactively was difficult for 22 Services[42], and this number declined to 17 one year later[43]. In this report, the number is 5[44]. The reason for this sharp decline can be largely attributed to the fact that the proactive moderation of China-based platforms is no longer in question, and to more Services disclosing the use of automated technologies to filter and block content.

Disclosures by Chinese Platforms

Finally, the first benchmarking report noted that the Chinese Services generally provided limited information about their content moderation practices and processes for enforcing their policies. In the second benchmarking report, TikTok stood out for being the first Chinese platform to disclose information on TVEC removals, although this information was only related to the international version of the app, and not to the local version Douyin. Other than this development, there were no changes in the remaining Chinese Services' disclosure approaches. Since the publication of the second benchmarking report, a lack of clear, accessible explanations of Chinese Services' content moderation practices remains the norm.

Both the first and second benchmarking reports explained that the Chinese Services' limited disclosures regarding content moderation and monitoring were likely the result of striking a balance between their obligation to comply with local laws and regulations and their need to keep their services attractive. The regulatory environment in China creates a system of intermediary liability under which online content sharing services have legal responsibility for content control (Knockel et al., 2018[53]), and the Chinese government has successively introduced stricter requirements in an effort to increase its control over Internet traffic and content. Companies must invest in staff and filtering technologies to moderate content and stay in compliance with governmental rules (Knockel et al., 2018[53]) such as the June 2017 cybersecurity law, which inter alia mandates the immediate removal of banned content and obliges Internet-based Services to assist security agencies with investigations (Creemers, 2018[54]). An admission of content monitoring in line with the prescriptions of the Chinese government could render Chinese Services unappealing on account of privacy and censorship concerns.

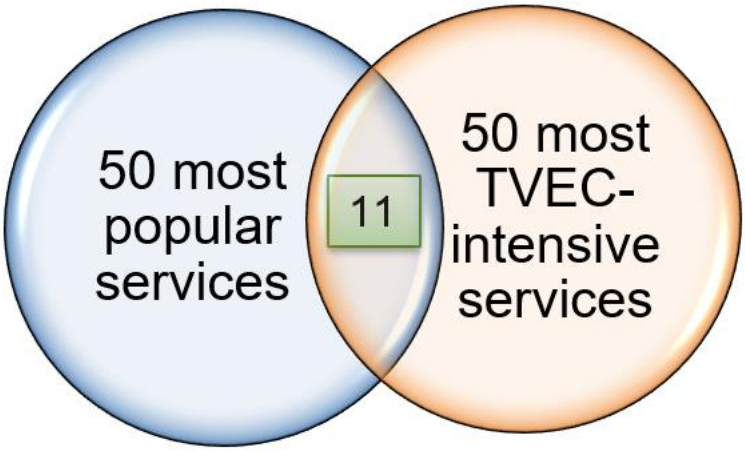
Although the Chinese Services endeavour to keep their content monitoring practices secret, multiple media reports and former employees' testimonies have confirmed these Services' collaboration with the Chinese government to filter content and censor speech. For example, Global Times, a Chinese state media outlet, reported that content-sharing services were expanding their human censor teams and developing artificial intelligence tools to review 'trillions of posts, voice messages, photos and videos every day', in an effort to make sure their content is in line with laws and regulations (Zhang, 2018[55]). State censorship authorities reportedly update a list of keywords and distribute it to platform operators on an ongoing basis (Wang, 2019[56]).

Academic research has shown that 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 and Shields Dobson, 2019[57]). Researchers from Citizenlab have empirically shown that WeChat implements different keyword-based and scanning tools to monitor content and improve China's surveillance mechanisms (Ruan, 2016[58]). Former employees of Chinese platforms increasingly come forward with testimony on how they developed tools and algorithms for content moderation in line with Chinese regulations, also noting that ByteDance, the controller of popular platforms such as Douyin, Toutiao, Xigua and Huoshan, employs a team of over 20 000 human moderators in China (Lu, 2021[59]).

On account of the abovementioned developments, the fact that the Chinese Services employ automated tools and human moderators to proactively and reactively moderate content is no longer in question. This fact has been reflected in some of this report's findings contained in this section (in particular, in the subsections on Staff member moderators, user-moderators and automated tools, and on Notifications, enforcement and appeal mechanisms and processes).

3. Commonalities, Developments and Trends in the Top-50 TVEC-intensive Services' Approaches to TVEC

Turning to the 50 services that are most widely used for accessing and disseminating TVEC, they are a largely different group from the 50 most popular services. Only 11 services appear on both lists.



The degree of elaboration and detail of TVEC-intensive Services' governing documents varies significantly

While some TVEC-intensive Services have clearly defined TVEC and their content moderation approaches (including notifications and appeal processes concerning moderation decisions), the majority of TVEC-intensive Services have overly broad or no terms of service whatsoever, especially those falling within the file-sharing and far-right-focused categories.

In particular, whereas some TVEC-intensive Services have clear prohibitions on TVEC, others are entirely silent in this regard.

Table 6 – TVEC definitions in TVEC-intensive Services' governing documents

Degree of elaboration of TVEC definitions	Mainstream (10)	File Sharing (15)	Far-right-focused (25)[45]
Governing documents that define terrorism, violent extremism and related concepts with sufficient detail to understand the scope of such terms, providing examples where appropriate	6[46]	0	0
Governing documents 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	1[47]	6[48]	5[49]
Governing documents that use broad and/or vague descriptions of prohibited conduct, which descriptions can be interpreted as supersets encompassing TVEC	3[50]	6[51]	9[52]
No content prohibition / Terms of Service	0	3[53]	10[54]

Further, as seen in Table 7 below, some TVEC-intensive Services endeavour to explain in detail their approach to content moderation, whilst others merely provide a contact form to report illegal content. A significant number of TVEC-intensive Services provide no information at all.

Table 7 – TVEC-intensive Services' content moderation approaches

Degree of elaboration of content moderation approaches	Mainstream (10)	File Sharing (15)	Far-right-focused (25)[55]
Content moderation approaches explained with good detail (e.g. a specific section in ToS or a blogpost explaining the service' overall approach, content moderation guidelines)	7[56]	6[57]	0
Content moderation approaches explained in broader / less detailed terms (e.g. "you can contact us via email at report@xxx.x and we'll review your complaint")	3[58]	2[59]	7[60]
Vague statements on content moderation (e.g. "we have the right but not the obligation to remove content…")	0	0	3[61]

Provision of contact form only	0	2[62]	2[63]
No information on content moderation available	0	5[64]	12[65]

In a similar vein, determining whether the TVEC-intensive Services have measures in place to ensure due process is not always possible. Some are explicit about their notification mechanisms and appeal processes, whilst others less so. In the majority of the cases, there is no available information in this regard.

Table 8 – TVEC-intensive Services' notifications and appeals mechanisms

Notifications / appeals in place	Mainstream (10)	File Sharing (15)	Far-right-focused (25)
TVEC-intensive Services that have mechanisms for notifying users in case of potential violations of their ToS and other governing documents	7[66]	6[67]	3[68]
TVEC-intensive Services that have appeal processes in place in respect of content moderation decisions and other measures applied under their governing documents	8[69]	6[70]	4[71]
No notifications and appeals specified / no information available	2[72]	9[73]	19[74]

Several factors may explain the variance in the degree of elaboration of the TVEC-intensive Services' governing documents and approaches to content moderation. First, the TVEC-intensive Services have markedly dissimilar financial resources. Nine of the ten TVEC-intensive Services falling within the mainstream category and two of the fifteen TVEC-intensive Services falling within the file-sharing category are included in the top-50 Services list (see Annex A) and generate substantial revenues from their large user bases. In fact, some of such Services' market capitalisation exceeds the gross domestic product of small countries (Access Now, 2020[60]). These TVEC-intensive Services tend to have carefully prepared governing documents and more sophisticated content moderation mechanisms in force. Conversely, many of the platforms and websites included in the file-sharing and far-right-focused categories are small, lack outside investment and generate limited revenues. Accordingly, they likely have limited resources to draw from and less capability to craft sound policies, processes and systems in support of content moderation efforts (Tech Against Terrorism, 2021[29]). Terrorist and violent extremist groups can thus avail themselves of the lack of content moderation policies and mechanisms in place to exploit these small platforms and websites. TVEC is likely to remain unmoderated there for longer, attracting more terrorists and violent extremists, who in turn disseminate more TVEC, thereby reinforcing TVEC prevalence on such platforms and websites.

Nonetheless, differences in financial power, though significant, only tell one part of the story. Many TVEC-intensive Services are self-declared champions of freedom of speech – e.g. BitChute[75], Rumble[76], Gab[77], Parler[78], Gettr[79], Redvoicemedia[80], Wimkin[81] and Xephula[82] – that contrast themselves with "Big Tech" firms and their perceived draconian content moderation practices. These "free speech platforms" typically pride themselves on their lax or non-existent content moderation efforts and convey the message that, save for activity such as child pornography, virtually everything is allowed on their platforms, regardless of how shockingly hateful and violent it may be. Thus, these particular TVEC-intensive Services cannot be said to be *exploited* by terrorist and violent extremist groups; rather, they purposefully *intend* to provide a space in

which extremist views and content can be expressed and disseminated without any fear of being banned or even criticised. In this setting, hate speech and calls to violence can proliferate in great volumes (Jasser and McSwiney, 2021[25]).

Worryingly, many TVEC-intensive Services are operated by terrorist and violent extremist organisations. Some video-sharing and news aggregator platforms like Brandnewtube, Worldtruthvideos and Mzwnews.com have been especially designed to cater to (mostly far-right) TVEC that has been removed from mainstream platforms, whereas websites like Nordfront.dk, Lookaheadamerica.org, Patriotfront.us, Vastarinta.com and Nordicresistancemovement.org openly espouse hateful ideologies and white supremacist views[83].

It follows from the above that there is no single solution to tackle the problem of TVEC online. As explained in the Introduction of this report, strong content moderation by mainstream platforms has caused a displacement effect whereby terrorist and violent extremist groups migrate elsewhere, flocking to lesser known platforms and websites. This has resulted in the exploitation of small platforms with limited ability or resources to address the problem. For these platforms, the creation of "a fund to finance and implement technical solutions to identify and manage the removal of [TVEC]" (Tech Against Terrorism, 2021[29]) is a sound proposal to curb the displacement effect. However, financial and technical support are bound to be fruitless in respect of platforms and websites that either have no intent to stop the spread of TVEC or actively strive to disseminate it. Given the multi-jurisdictional nature of online domains, a clampdown on these TVEC-intensive Services will require international agreement and coordination on multiple levels – including what fundamental freedoms to prioritise, laws to enforce and technical measures to implement.

Transparency Reporting on TVEC by TVEC-intensive Services is uncommon

Transparency reporting on policies and content moderation efforts concerning TVEC is rather unusual amongst the TVEC-intensive Services (eight out of 50). The majority of TVEC-intensive Services that issue transparency reports on TVEC are in the mainstream category and are all on the top-50 most popular Services list (see Annex A), as well, for which reason the content of their transparency reports is discussed in Section 3 above[84]. Of the 50 TVEC-intensive Services that are not included in the top-50 (most popular) Services list, only two issue transparency reports expressly addressing TVEC.

Table 9 – TVEC-intensive Services that issue transparency reports with information on TVEC

Mainstream	File-sharing	Far-right-focused
YouTube Twitter Facebook Instagram TikTok Discord	Justpaste.it Mega.nz	None

Although the number of TVEC-intensive Services joining the select group of 15 Services that issue transparency reports on TVEC is low, the engagement of these two platforms in this practice is particularly noteworthy due to the type of functionality they provide, and in the case of Justpaste.it, due to its size.

Mega is a privacy-by-design cloud storage and communication (chat) platform which relies on end-to-end encryption and provides zero-knowledge privacy[85] (Mega.nz, 2021[61]). Its security- and privacy-focused design is its main selling point. Thus, Mega explains that it does not scan stored files and then compare hash values to industry hash sets because files are encrypted on clients' devices before being uploaded and the stored encrypted file has a different hash to the original file[86]. Nonetheless, Mega clarifies that it receives a few reports of child sexual abuse material (CSAM) links from international NGOs (such as reporting hotlines) and from law enforcement agencies, but most are submitted by private individuals who have noticed the links, with an associated description, being openly shared on public forums. In turn, most reports of violent extremism links are provided by an NGO which monitors public websites. Mega observes that anyone with the link, including the decryption key, can download the content so Mega can proceed to immediately disable the link and close the infringer's account (Mega.nz, 2021[62]). Mega is hardly a novice in transparency reporting. It has published seven transparency reports since it commenced operations in 2013, with the explicit intent to "provide transparency for users, regulatory bodies and suppliers, as to Mega's operating processes relating to privacy and to statutory compliance" (Mega.nz, 2021[62]). In its 2021 transparency report, Mega reported the following: 1) the total number of accounts closed for sharing objectionable content (which includes violent extremism); 2) the number of reported violent extremism links that were disabled (numbers reported cover Q4 2019, 2020 and 2021); 3) the number of violent extremism link reports broken down by sources - NGOs, law enforcement, individuals, industry (numbers reported cover Q4 2019, 2020 and 2021); and 4) the number of warrants Mega received during the reporting period for violent extremism, indicating the originating country and the outcome of the warrant (e.g. metadata supplied).

The case of Mega shows that privacy and safety are not inherently at odds with one another because it uses end-to-end encryption and still publishes useful transparency reports on content moderation, including TVEC. Even if technically prevented from monitoring and detecting objectionable or otherwise illegal content, such content can still be brought to the attention of platform operators, which increasingly face TVEC removal obligations (see Section 4 of this report) and therefore are bound to act on TVEC reports. To shed light on the overall TVEC online landscape, reinforce user trust, enable accountability and facilitate informed policy-making in the fight against TVEC, it is essential that services offering encrypted or otherwise security- and privacy-driven functionalities follow Mega's example.

Justpaste.it is a text- and images-sharing platform which caters to privacy-sensitive users. It is anonymous by default, which means that users do not have to create an account to publish something, so there is no need to disclose one's name, email or home address (Justepaste.it, 2018[63]). This feature was exploited by IS, which disseminated TVEC on Justpaste.it (Ilinsky, 2019[64]). Justpaste.it quickly collaborated with law enforcement to remove the TVEC uploaded on the platform, and since then it became a member of GIFCT, relying on its Hash Sharing database to keep the platform TVEC-free. Thus, Justpaste.it uses anonymous identifiers of content that were shared by other companies to detect terrorist content on its platform[87]. To reinforce user trust and its commitment to privacy, since 2019 Justpaste.it publishes transparency reports with TVEC-related information. In its last transparency report (covering 2021), Justpaste.it reported the number of requests received from governments and law enforcement agencies regarding illegal content (broken down by EU requests, UK requests, Russian requests and Turkish requests), the percentage of such requests that related to terrorist content, and the percentage of terrorist content reports on which Justpaste.it took action and blocked the terrorist content.

Justpaste.it's quick collaboration with law enforcement, its participation in industry-led initiatives to combat TVEC such as the GIFCT, and its commitment to transparency is remarkable given that Justpaste.it is "a one man show" with very limited monetisation streams: the company's founder, Mariusz Zurawek, is its

only employee, and Justpaste.it makes all of its revenues from ads run by Google AdSense (Ilinsky, 2019[64]). Hence, whereas the time and costs involved in devising sound content moderation policies and procedures and producing transparency reports must be acknowledged as a challenge for small-sized platforms, the example of Justpaste.it shows that it can be met.

4. TVEC-related Laws and Regulations that Have Been Enacted or Are under Consideration

Governments initially tended to rely on self-regulation and voluntary pledges to nudge Internet-based services into taking a more aggressive stance on TVEC online. However, industry efforts to counter TVEC have been perceived as inadequate, thus triggering a shift from the self-regulatory model towards increased government intervention (Gorwa, 2019[65]). In particular, a growing number of governments are proposing and enacting laws and regulations to thwart the spread of TVEC online. This Section provides an overview of laws and regulations addressing TVEC that have been recently enacted, or that are currently under consideration, in OECD jurisdictions.

Australia

The Online Safety Act 2021 (the Online Safety Act), which came into force in January 2022, reforms and expands existing online safety regulations in Australia, consolidating and modernising separate regulatory regimes. This includes updating the schemes on cyber-bullying of children, image-based abuse (the threatened or actual non-consensual distribution of intimate images) and illegal and restricted online content (such as TVEC), and introduces the adult cyber abuse scheme.

The Online Safety Act gives the eSafety Commissioner specific powers and functions in relation to 'class 1 material'. TVEC will ordinarily fall within the definition of class 1 material, which is material that is, or is likely to be, refused classification under the National Classification code, and includes material that directly or indirectly counsels, promotes, encourages, urges, provides instruction on or praises the doing of a terrorist act (see section 106 of the Online Safety Act and section 9A of the *Classification (Publications, Films and Computer Games) act 1995;* A 'terrorist act' is defined in section 100.1 of the *Criminal Code*).

Regarding TVEC online, the Online Safety Act provides the eSafety Commissioner with powers to:

- Investigate, in response to a complaint from the public or on the Commissioner's own initiative, whether Australians can access certain material, including class 1 material such as TVEC;
- Take action to require the removal of certain material, including class 1 material such as TVEC, from particular types of digital services;

- Request or require Internet service providers to block sites for short periods of time in online crisis events involving the distribution of material that promotes, incites, instructs in or depicts 'abhorrent violent conduct', including terrorist acts;
- Register industry-developed mandatory codes to address class 1 material such as TVEC on a systemic basis; and
- Require online service providers to report on how they are meeting a set of Basic Online Safety Expectations, including an expectation to minimise the provision of TVEC.

Online Content Scheme

The eSafety Commissioner may investigate whether end-users in Australia can access class 1 material, such as TVEC, provided on a social media service, designated Internet service or relevant electronic service.

The eSafety Commissioner has the power to issue removal notices to services anywhere in the world that provide and/or host class 1 material. Non-compliant services may be subject to civil penalties.

The eSafety Commissioner also has other administrative powers which can be used to address TVEC online. For example, if the eSafety Commissioner is satisfied that there were two or more times during the previous 12 months when end-users in Australia could use a service to download an app that facilitates the posting of class 1 material, then the Commissioner can give a notice to require an app distribution service to cease enabling end-users in Australia to download an app that facilitates the posting of class 1 material on a social media service, relevant electronic service, or designated Internet service in particular circumstances. Similarly, the eSafety Commissioner may give a notice to the provider of an Internet search engine requiring it to cease providing a link to class 1 material if the Commissioner is satisfied that there were two or more times during the previous 12 months when end-users could access class 1 material using a link provided by the service.

ISP Blocking

The Online Safety Act enables the eSafety Commissioner to give Internet service providers ablocking request or a blocking notice when material that promotes, incites, instructs in or depicts abhorrent violent conduct can be accessed using a service supplied by an Internet service provider. Abhorrent violent conduct is defined to mean murder or attempted murder, a terrorist act, torture, rape or kidnapping. The eSafety Commissioner will use these powers only in situations where the Commissioner has declared an online crisis event (Parliament of the Commonwealth of Australia, House of Representatives, 2021[66]).

The power to request or require blocking enables the eSafety Commissioner to direct Internet service providers to block domains and websites containing TVEC for time limited periods. The blocking request or notice can include a number of specified steps to disable access to the material, including steps to block domain names that provide access to the material, steps to block URLs that provide access to the material, and steps to block IP addresses that provide access to the material.

To issue blocking requests or notices, the eSafety Commissioner must determine that the availability of the material is likely to cause significant harm to the Australian community. This determination hinges on the nature of the material, the number of end-users who are likely to access the material, and other matters the eSafety Commissioner may deem relevant. The Commissioner must also have regard to whether any other powers conferred to them could be used to minimise the likelihood that the availability of the material online could cause significant harm to the Australian community.

Industry Codes or Industry Standards

The Online Safety Act provides for industry bodies or associations to develop new codes to regulate certain types of harmful online material, including material that advocates committing a terrorist act, and for the eSafety Commissioner to register the codes. The proposed industry codes cover eight key sections of the online industry, including providers of social media, messaging, search engine and app distribution services, as well as Internet and hosting service providers, manufacturers and suppliers of equipment used to access online services and those that install and maintain equipment.

In September 2021, the eSafety Commissioner released a position paper (eSafety Commissioner[67]) to help the online industry develop codes. The paper set out 11 policy positions regarding the substance, design, development and administration of industry codes, as well as the Commissioner's preferred outcomes-based model for the codes.

On 11 April 2022, the eSafety Commissioner issued notices formally requesting the development of phase 1 of the industry codes. These were issued to six industry associations that formed a steering group to oversee codes development for the eight industry sections outlined under the Online Safety Act. These industry associations are:

- Australian Mobile Telecommunications Association
- BSA – The Software Alliance
- Communications Alliance
- Consumer Electronics Suppliers' Association
- Digital Industry Group Inc
- Interactive Games and Entertainment Association

Public and industry consultation on the draft codes, which were released on 1 September 2022, is the responsibility of the industry associations. Industry's draft codes, and information about how to make a submission, can be found at onlinesafety.org.au.

Draft industry codes for each of the eight sections of the online industry identified in the OSA will be submitted to eSafety in late 2022. If an industry code meets the statutory requirements, the eSafety Commissioner will register that industry code and it will be enforceable in respect of the activities of those participants whose activities are covered by that industry code. If an industry code provided for registration does not meet the statutory requirements, the eSafety Commissioner is able to determine an industry standard for that section of the online industry.

Basic Online Safety Expectations

The Online Safety Act also allows for the setting of basic online safety expectations by the relevant Minister. These expectations, registered in January 2022, include that a provider of a social media service, designated Internet service, or relevant electronic service should:

- take reasonable steps to minimise the extent to which material that promotes, incites, instructs in or depicts abhorrent violent conduct can be accessed on that service;
- ensure that they have clear and readily identifiable mechanisms that enable end-users to report, and make complaints about, material that promotes, incites, instructs in or depicts abhorrent violent material and breaches the service's terms of use; and
- disclose, upon request, specific information about online harms to eSafety.

The expectations themselves are not enforceable. However, the Act provides the eSafety Commissioner with powers to require services to report on their implementation of the expectations, in the manner and

form specified. These notices are enforceable, and backed by civil penalties. Reporting notices are specific to the provider, although multiple notices can be issued. Notices can be for:

- non-periodic reporting
- periodic reporting over a specified time frame of between six to 24 months.

The eSafety Commissioner can also make reporting determinations – a legislative instrument – requiring periodic or non-periodic reporting for a specified class of services. Like the reporting notices, these are enforceable and backed by civil penalties for failure to report.

Finally, the eSafety Commissioner can issue statements regarding whether providers are meeting the expectations.

The eSafety Commissioner issued the first non-periodic reporting notices (eSafety Commissioner, 2022[68]) in August 2022 requiring information about how providers are meeting certain expectations relating to child sexual exploitation and abuse.

Canada

Canada's Digital Charter (2019) outlines Canada's approach to internet-based technologies and online space governance (Government of Canada, 2019[69]). Its 9th principle underscores that Canadian citizens should expect that digital platforms will not foster or disseminate hate, violent extremism or criminal content.

In mandate letters issued by Canada's Prime Minister dated 16 December 2021, sent to the Minister of Justice and the Minister of Canadian Heritage, the introduction of legislation was designated as crucial to combat serious forms of harmful online content and hold social media platforms and other services accountable for the content they host, including by strengthening the *Canadian Human Rights Act* and the *Criminal Code* to combat online hate and reinforce hate speech provisions (Trudeau, 2021[70]) (Trudeau, 2021[71]).

On 30 March 2022 the Government of Canada established an expert advisory group on online safety, mandated to provide the Minister of Canadian Heritage with advice on how to design a legislative and regulatory framework to address harmful content online. Work with the expert group is underway. Summaries of each session can be found on the Canadian Heritage website[88].

European Union

During plenary of the European Parliament on 28 April 2021, the "Regulation to address the dissemination of terrorist content online" was finally adopted. The Regulation imposes on hosting service providers established in the European Union the obligation to address the misuse of their platforms by terrorists. National competent authorities are empowered 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 are 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.

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

- are bound to set out in their terms and conditions their policy for addressing the dissemination of terrorist content; and
- 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.

Further, with an aim to clarify the responsibilities and strengthen the accountability of services that intermediate content, the European Commission adopted in December 2020 a proposal for the Digital Services Act (DSA)[89]. The DSA imposes new obligations on digital service providers centred around four main principles:

- Transparency: All digital service providers without an establishment in the EU must appoint a legal representative in a Member State where they offer services. They must publish clear, comprehensible and detailed annual reports on content moderation (with additional information required for online platforms and "very large online platforms", VLOPs). Online platforms must also ensure that traders provide sufficient information to the platform and display trader information to users. Online platforms must provide transparency on advertisements and on the algorithms used to display them (with additional requirements for VLOPs). VLOPs must also publish information on their use of recommender systems.
- Empowering users: All digital service providers must include information on any content restrictions that they impose in their terms and conditions. Providers of hosting services must set up a notice mechanism for users to report illegal content and they must give a statement of reasons when they remove or disable access to specific content. Online platforms must provide content dispute resolution mechanisms enabling users to appeal their decisions.
- Risk management: Online platforms must take measures to protect their systems against misuse, including obligations to remove illegal goods, services or content. Online platforms must inform the relevant authorities if they suspect a serious criminal offence involving a threat to the life or safety of persons. VLOPs must also take steps to manage systemic risks, including annual risk assessments, risk mitigation measures, annual independent audits and appointing compliance officers.
- Industry co-operation: The European Commission will support and promote the development of voluntary industry standards, codes of conduct and crisis protocols on certain aspects of online businesses.

VLOPs have specific obligations in relation to certain types of harmful content. They must assess, and take steps to mitigate, systemic risk to users of their service in relation to:

- the dissemination of illegal content;
- negative effects for the exercise of fundamental rights, for example the right to private and family life, freedom of expression etc.; and
- intentional manipulation of their services with an actual or foreseeable effect on public health, minors, civic discourse, electoral process or public security.

Measures that may be needed to address these risks could include adapting content moderation or recommender systems, discontinuing advertising revenue for specific content, and improving the visibility of authoritative information sources.

The concept of "illegal content" is defined broadly and refers to information that under applicable law (EU and/or relevant Member State) is either itself illegal, such as illegal hate speech or terrorist content, or relates to activities that are illegal, such as the sharing of images depicting child sexual abuse.

Each Member State will be required to appoint a Digital Services Coordinator (DSC) to enforce the DSA. If a DSC finds that a digital service provider has breached its obligations, it will have the power to:

- order the cessation of infringements;
- impose interim measures; and
- impose fines of up to 6% of the infringer's global annual turnover, or periodic penalty payments of up to 5% of the infringer's average global daily turnover.

In cases concerning VLOPs, the issue can be escalated to the Commission.

France

Law n°2020-766 of 24 June 2020 on hate speech on the Internet, also called the "Avia Law" named after the law's main sponsor), came into force on 26 June 2020. Its provisions modify Law n°2004-575 of 21 June 2004 on Confidence in the Digital Economy ("LCEN"), which largely mirrors the provisions of the EU eCommerce Directive.

The Avia Law was intended to be much broader than its current scope. It originally included provisions similar to those found in the NetzDG in Germany. However, many of those provisions were quashed by the French Constitutional Court on 18 June 2020. The court held that a proposed requirement that platforms remove "manifestly" illegal content within 24 hours was incompatible with the right to freedom of expression, given the risk that platforms would "over-block" to avoid enforcement action.

The LCEN requires online communications services and social media platforms to:

- set up an easily accessible system to allow users to report hate speech;
- publish details of the resources they devote to tackling hate speech on their platforms;
- remove child sexual abuse or terrorist content within 24 hours of being notified of the material by the general directorate of the national police; and
- promptly inform the competent public authorities of harmful content reported to them. The law does not define "promptly", but case law suggests that a delay of five days is too long. The platform must also provide any data they hold which would help to identify the user who posted the content.

The LCEN provides an exhaustive list of "hate speech content", i.e. anything that breaches the French Criminal Code or the French Law on the Freedom of the Press of 29 July 1881. This includes:

- sexual harassment;
- provoking hatred or violence against a person based on their gender, sexual orientation, disability, race, or religion; and
- directly provoking or condoning terrorist acts.

The Avia Law also created a research body to monitor and analyse the development of online hate speech. The courts have broad powers to enforce the regime, including orders to block access to certain websites. The French general directorate of the national police has the power to demand the removal of child sexual abuse and terrorist content.

The Avia Law introduced significant penalties in case of non-compliance:

- for individuals, including directors and senior employees of service providers, fines of up to EUR 250 000 and one-year imprisonment; and
- for companies, fines of up to EUR 1.25 million and a prohibition preventing them from carrying out their activity for five years.

Lastly, the French government introduced in December 2020 the Endorsement of Respect for the Principles of the Republic and Counter Separatism Bill (commonly known as the "Bill against separatism"). It is deemed a key-pillar of the government's strategy to counter Islamist radicalisation and terrorism. The Bill was approved after the first reading by the French parliament and senate, but it must go through another reading before being passed.

The Bill punishes:

- the malicious sharing of personal information online that endangers the life of others;
- those who directly incite, legitimise or praise terrorism, with a 7-year prison sentence and up to EUR 100 000 in fines. This applies to content shared on messaging platforms;
- individuals who deliberately seek to circumvent moderation techniques used to counter and delete banned content.

The Bill also creates new obligations for online platforms, notably with regard to disclosing information about their algorithms and content moderation process.

Germany

Germany's Act to Improve Enforcement of the Law in Social Networks ("NetzDG") came into force in 2017. It was last amended in April 2021 by the Act to Combat Right-Wing Extremism and Hate Crime, with a view to tackling hate crime and other harmful content on social networks.

Under the NetzDG, "providers of social networks" are service providers who, for profit, operate internet platforms designed to allow users to share content, regardless of where they are established. "Content" includes own and third-party content, such as images, video and text.

Providers of social networks must implement effective control mechanisms to filter, block or take down unlawful content on their platforms. This means:

- having an effective and transparent system for managing user reports;
- offering users an easily accessible way to report any unlawful content;
- reviewing any reported content expeditiously. "Manifestly unlawful" content must be blocked or removed within 24 hours. Any other unlawful content must be blocked or removed within seven days
- reporting harmful content to the Federal Criminal Police Office when the content expresses certain criminal expressions. The list of expressions is included in 3a (2) of the NetzDG. It includes:
  - child pornography
  - dissemination of propaganda and symbols from anti-constitutional organisations
  - preparation of violent action against the state
  - education and support of criminal or terrorist associations
  - incitement to hatred
  - representation of violence
- complying with a range of reporting and transparency obligations, such as:

- providing information on content moderation procedures.
- detailing the results of their use of automated methods for detecting illegal content, and
- clarify whether the provider has given access to their data to independent researchers
- service providers which receive more than 100 reports about unlawful content per calendar year must publish bi-annual reports on their reports handling process.

Non-compliance with the NetzDG's obligations may lead to fines amounting to:

- up to EUR 5 million against the provider's representatives (for example, the managing director, or the owner), where they are responsible for the non-compliance; and
- up to EUR 50 million against the legal person or association of persons operating the platform. Because the fine should exceed the economic advantage of the platform, the fine may also exceed the EUR 50 million cap in specific cases where the platform's economic advantage is higher than EUR 50 million.

As noted above, the NetzDG was amended in April 2021 by the Act to Combat Right-Wing Extremism and Hate Crime, simplifying the prosecution of right-wing extremism and hate crime offences. The amendment creates an obligation for platforms to report certain types of unlawful content, such as online threats, and other information such as the IP address of the respective user to the Federal Criminal Police Office.

The NetzDG was further amended by an Act intended to strengthen the rights of users of social networks by making reporting channels more user-friendly, creating more transparency by expanding the scope of the biannual reports, and creating a right for both the users and the individuals/groups who reported the content to appeal against decisions of the platform not to block or remove reported content.

Ireland

Ireland does not currently have specific legislation governing online harms. However, in January 2020, the Online Safety and Media Regulation Bill was proposed. The Bill is the first piece of legislation that deals with video regulation of video-sharing platforms, including YouTube. The Bill also updates the way in which television broadcasting services and video on-demand services are regulated, and aims to ensure greater regulatory alignment between traditional linear TV and video on-demand services, such as RTÉ Player and Apple TV.

The Bill applies to "relevant online services" (any information society service established in Ireland that allows a user to disseminate or access user-generated content); and "designated online services" (any relevant online service that has been designated as such by the Media Commission). This means that a wide range of different organisations may come within the scope of the new law, including video service providers, social media providers, e-commerce services, online search engines and Internet service providers.

The Bill proposes the establishment of a Media Commission, including an Online Safety Commissioner. This new body will replace the Broadcasting Authority of Ireland and will also be responsible for the regulation of on-demand services, including radio, television, and video-on-demand services.

Under the Bill, online services must to comply with Online Safety Codes prepared by the Media Commission. These codes may require providers to take actions like:

- minimising the availability of harmful online content;
- implementing measures in relation to commercial communications available on their services. putting in place mechanisms to handle user complaints and issues;

- carrying out risk and impact assessments in relation to the availability of harmful online content on their services; and
- reporting obligations regarding compliance with the Online Safety Codes.

The Bill sets out four categories of harmful online content:

- material which it is a criminal offence to disseminate under Irish or EU law (for example, child sexual abuse material or terrorist content);
- cyber-bullying material;
- material encouraging or promoting eating disorders; and
- material encouraging or promoting self-harm or suicide. The Media Commission will be able to include or exclude
- other categories of harm.

The Media Commission will have the power to:

- designate online services for regulation;
- prepare, monitor and conduct investigations into compliance with the Online Safety Codes;
- audit any complaints or issues handling processes;
- operate a "super complaints" system, where nominated bodies can bring systemic issues to the Media Commission's attention; and
- direct online services to make changes to their systems, processes, policies and design.

Also, the Media Commission will have a broad range of enforcement powers, including:

- issuing information requests, compliance notices, and warning notices mandating compliance (which can be published). In Ireland, failure to comply with an information request or warning notice is a criminal offence, punishable by a fine of up to EUR 5,000 and/or 1-year imprisonment (on summary conviction).
- pursuing civil sanctions, including administrative fines of up to EUR 20 million or 10% of relevant turnover (whichever is higher) for the preceding financial year, issuing orders compelling compliance with warning notices, or requiring internet service providers to block access to the offending online service in Ireland.

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[90]. 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[91] (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.

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.

Türkiye

In Türkiye, Additional Article 4 of Law No. 5651 (known as the Social Media Bill) passed in July 2020 by the Turkish parliament and entering into force on 1 October 2020, implemented a wide range of new requirements and steep penalties for social media companies. Under the new Social Media Bill:

- Social network providers with over a million daily users in Türkiye are required to establish a formal presence in-country by assigning a natural or legal person representative. If foreign social network providers with over a million daily users in Türkiye do not comply with the representative requirement, gradually authorities are able to impose gradual fines such as administrative monetary penalty, advertising band and bandwidth throttling up to 90%.
- All social network providers have to respond to user complaints about content that 'violate personal and privacy rights' within 48 hours, if not face a fine of 5 million Turkish Liras.

- All social network providers are required to remove or block access to content related to crimes listed in the Turkish Criminal Code (as determined by a judge or a court) within 4 hours. If social network providers do not comply with this order, they will face fines of up to 1 million Turkish Liras. Also, social network providers face access blocking regarding the related content (as determined by a judge or a court) on their platforms.
- All social network providers with over a million daily users in Türkiye are required to prepare a report every six months in Turkish, containing statistical and categorical data regarding (i) user complaints about contents that violate personal and privacy rights and (ii) their compliance with the court decisions about content removal or access blocking. This report also has to be publicly accessible on the social network provider's platform. If social network providers do not comply with these obligations, they will face a fine of 10 million Turkish Liras (around 1 million USD).
- All social network providers with over a million daily users in Türkiye are required to take "necessary measures" to host the data of Turkish users in Türkiye – that is, they have data localisation obligations.

In Türkiye, Law No. 6112 regulates video on-demand platforms. Under this law, video on-demand platforms must get a licence from Turkish authorities. Turkish authorities have the power to remove inappropriate content and take actions in respect of content which may be harmful to juveniles.

United Kingdom

In 2019, the United Kingdom outlined a comprehensive plan for online regulation in the Online Harms White Paper, aiming to make the United Kingdom "the safest place in the world to be online" (HM Government, 2019[72]). The Paper aimed to counter various online harms ranging from cyberbullying to terrorist content. The 2019 Paper was followed in 2020 by the adoption of The Interim Code of Practice on Terrorist Content and Activity Online and The Interim Approach for regulating video-sharing platforms. In May 2021, the draft Online Safety Bill was finally published, and is currently undergoing a second reading in the House of Commons[92].

The Bill requires those services falling within its scope to

- assess their user base and the risks of harm to those users present on the service;
- take steps to mitigate and manage the risks of harm to individuals arising from illegal content and activity, and (for services likely to be accessed by children) content and activity that is harmful to children;
- put in place systems and processes which allow users and affected persons to report specified types of content and activity to the service provider;
- establish a transparent and easy to use complaints procedure which allows for complaints of specified types to be made;
- have regard to the importance of protecting users' legal rights to freedom of expression and protecting users from a breach of a legal right to privacy when implementing safety policies and procedures; and
- put in place systems and processes designed to ensure that detected but unreported child sexual exploitation and abuse (CSEA) content is reported to the National Crime Agency (NCA).

The harm caused by the illegal or otherwise harmful content is either physical or psychological. Illegal content includes content that amount to terrorist, CSEA and other criminal offences (set out in Schedule 7) under UK law, whereas harmful content comprises a range of content the description of which is to be designated in regulations by the Secretary of State.

Ofcom, the communications industry regulator, will be the regulator for the online safety regime. If Ofcom finds that a platform has failed to comply with its regulatory obligations under the new regime, it will have a broad range of enforcement options, including the power to:

- issue a fine of up to the greater of GBP 18 million or 10% of the platform's annual turnover;
- require third parties to withdraw access to key services that make it less commercially viable for the platform to operate within the United Kingdom; and
- require key internet infrastructure service providers to take steps to block a platform's services from being accessible in the United Kingdom, for example ISP blocking.

Companies falling within the scope of the new regime 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. To uphold their duty of care, regulatees will be bound to *inter alia*:

- update their terms of Service to explicitly mention which content they deem appropriate (or inappropriate) on their platforms;
- produce annual transparency reports;
- introduce an easy-to-access user complaints system; and
- take appropriate action in response to complaints of a relevant kind.

United States

In the United States, there is no legislation that requires platforms to take measures in respect of harmful content online and TVEC. Indeed, section 230 of the Communications Act of 1934 as amended by the Telecommunications Act of 1996 provides that "no provider or user of an interactive computer service shall be treated as the publisher or speaker of any of the information provided by another information content provider". This provision gives online platforms and Internet service providers broad immunity from liability for user-generated content on their platforms.

Section 230 also protects platforms where they voluntarily take steps in "good faith" to moderate user-generated content, by ensuring they will not be held liable for their moderation decisions. This is intended to encourage platforms to engage in content moderation without fear of being held liable for these moderation decisions. In 2018, Section 230 was amended by the Stop Enabling Sex Traffickers Act (FOSTA-SESTA) to require the removal of material violating federal and state sex trafficking laws.

The United States' approach to online content regulation is influenced by its focus on freedom of speech, as set out in the First Amendment, 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.

Annex A - Global Top-50 Most Popular Online Content-Sharing Services

Rank	Name of service (parent company)	Monthly active users (*a*) or unique visitors (*b*) (millions)	Type of service	Issues TVEC transparency reports	Provided feedback / comments on its profile
1	Facebook (Facebook, Inc.)	2,853*(a)* (as of July 2021) (Datareportal, 2021[73])	Social networking platform	Y	N
2	YouTube (Alphabet, Inc.)	2,291*(a)* (as of July 2021) (Datareportal, 2021[73])	Video streaming platform	Y	N
3	Zoom (Zoom Video Communications, Inc.)	2,100*(b)* (as of June 2021) (Statista, 2021[74])	Video chat and voice calls app	Y	Y
4	WhatsApp (Facebook, Inc.)	2,000*(a)* (as of July 2021) (Datareportal, 2021[73])	Messaging app	N	N
5	iMessage/FaceTime (Apple, Inc)	1,650*(a)* (as of January 2021) (Kastrenakes, 2021[75]))	Messaging and video chat apps	N	N
6	Instagram (Facebook, Inc.)	1,386*(a)* (as of July 2021) (Datareportal, 2021[73])	Social networking platform	Y	N
7	Facebook Messenger (Facebook, Inc.)	1,300*(a)* (as of July 2021) (Datareportal, 2021[73])	Messaging app	Y (included in Facebook's)	N

8	Weixin/WeChat (Tencent Holdings Ltd.)	1,242*(a)* (as of July 2021) (Datareportal, 2021[73])	Social networking/content sharing/messaging platform	N	N
9	Viber (Rakuten, Inc.)	820*(a)* (as of January 2021) (99 Firms, 2021[76])	Messaging app	N	Y
10	Tik Tok (ByteDance Technology Co.)	732*(a)* (as of July 2021) (Datareportal, 2021[73])	Short video app	Y	Y
11	QQ (Tencent Holdings Ltd.)	606*(a)* (as of July 2021) (Datareportal, 2021[73]) as of July 2021	Instant messaging and web portal site	N	N
12	Youku Tudou (Alibaba Group Holding Limited)	600*(a)* (as of January 2021) (V-click Technology, 2021[77])	Video streaming platform (user-generated and syndicated content)	N	N
13	Telegram (Telegram Messenger LLP)	550*(a)* (as of July 2021) (Datareportal, 2021[73])	Messaging app	N	N
14	QZone (Tencent Holdings Ltd.)	548*(a)* (as of January 2021) (Warner, 2021[78])	Social networking platform	N	N
15	Weibo (Sina Corp.)	530*(a)* (as of July 2021) (Datareportal, 2021[73])	Social networking platform	N	N
16	Snapchat (Snap, Inc.)	514*(a)* (as of July 2021) (Datareportal, 2021[73])	Social networking platform	Y	Y
17	Kuaishou (Beijing Kuaishou Technology Co., Ltd)	481*(a)* (as of July 2021) (Datareportal, 2021[73])	Short video app	N	N
18	iQIYI (Baidu, Inc.)	480*(a)* (as of January 2020 (Statista, 2021[79])	Video streaming platform (user-generated and syndicated content)	N	N
19	Pinterest (Pinterest, Inc.)	478*(a)* (as of July 2021) (Datareportal, 2021[73])	Social networking platform	Y	N
20	Reddit (Reddit, Inc.)	430*(a)* (as of July 2021)	Social news aggregation, web	Y	Y

		(Datareportal, 2021[73])	content ranking and discussion website		
21	Twitter (Twitter, Inc.)	397*(a)* (as of July 2021) (Datareportal, 2021[73])	Short messages-focused social networking platform	Y	N
22	Tumblr (Automattic, Inc.)	327*(a)* (as of January 2021) (Finances Online, 2021[80])	Microblogging and social networking platform	N	Y
23	LinkedIn (Microsoft, Inc.)	310*(a)* (as of January 2021) (99 Firms, 2021[81])	Jobs-focused social networking platform	N	Y
24	Douban (Information Technology Company, Inc.)	300*(a)* (as of July 2021) (Marketing to China, 2021[82])	Social networking platform	N	N
25	Baidu Tieba (Baidu, Inc.)	300*(a)* (as of January 2021) (Marketing to China, 2021[83])	Online communications platform	N	N
26	Quora (Quora, Inc.)	300*(a)* (as of July 2021) (Datareportal, 2021[73])	Question-and-answer website	N	N
27	Teams (Microsoft, Inc.)	250*(a)* (as of July 2021) (Techcircle, 2021[84])	Online collaboration platform	Y	Y
28	IMO (PageBites, Inc.)	212*(a)* (as of January 2020) (Smith, 2021[85])	Video chat and voice calls app	N	N
29	Ask.fm (IAC [InterActiveCorp])	180*(b)* (as of January 2021) (Ask.fm, 2021[86])	Social networking platform	N	N
30	Vimeo (Vimeo, Inc.)	170*(a)* (August 2020) (Startup Talky, 2020[87])	Video streaming app	N	Y
31	Medium (A Medium Corporation.)	170*(b)* (as of January 2021) (Willens, 2021[88])	Online publishing platform	N	N
32	LINE (Line Corporation)	167*(a)* (as of July 2020) (Line Corporation, 2020[89])	Messaging app	N	N
33	PicsArt (PicsArt, Inc.)	150*(a)* (as of February 2020) (Sensor Tower, 2020[90])	Photo and video app	N	N

34	Discord (Discord, Inc.)	150*(a)* (as of January 2021) (Discord, 2021[91])	Chat platform	Y	N
35	Twitch (Amazon.com, Inc.)	140*(a)* (as of January 2021) (Dean, 2021[92])	Livestreaming platform	Y	N
36	Likee (BIGO Technology PTE. LTD.)	115*(a)* (as of January 2021) (JOYY Inc., 2021[93])	Streaming platform	N	N
37	Skype (Microsoft, Inc.)	100*(a)* (as of March 2020) (Lardinois, 2020[94])	Video chat and voice calls app	Y	Y
38	VK (Mail.Ru Group)	97*(a)* (as of January 2021) (Mail.ru, 2021[95])	Social networking platform	N	N
39	Xigua Video (ByteDance Technology Co.)	85*(a)* (as of May 2021) (Statista, 2021[96])	Short video streaming app	N	N
40	Odnoklassniki (Mail.Ru Group)	71*(b)* (as of January 2021) (Mail.ru, 2021[95])	Social networking platform	N	N
41	Flickr (SmugMug, Inc.)	60*(b)* (as of January 2021) (Flickr, 2021[97])	Image and video hosting service	N	N
42	Huoshan (ByteDance Technology Co.)	59*(a)* (as of May 2021) (Statista, 2021[96])	Short video streaming app	N	N
43	KaKao Talk (Daum Kakao Corporation)	52*(a)* (as of December) (Statista, 2021[98])	Messaging app	N	Y
44	Smule (Smule, Inc.)	50*(a)* (as of September 2020) (Audiens, 2020[99])	User-generated music-video sharing platform	N	N
45	Deviantart (DeviantArt, Inc.)	45*(b)* (as of January 2021) (DeviantArt, 2021[100])	Online artwork, videography and photography 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.)	35.69% (as of January 2021) (Datanyze, 2021[101])	Cloud-based file sharing	N	N
Dropbox (Dropbox, Inc.)	20.42% (as January 2021) (Datanyze, 2021[101])	Cloud-based file sharing	N	N
Microsoft OneDrive (Microsoft, Inc.)	13.66% (as of January 2021) (Datanyze, 2021[101])	Cloud-based file sharing	Y	Y

Name of service (parent company)	Indicative Global Market Share or monthly average unique devices (millions)	Type of market/service	Transparency report on terrorist/violent extremist content	Provided feedback / comments on its profile
Wordpress.com (Automattic, Inc.)	62% (as of January 2021) (Envisage Digital, 2021[102])	Content management system	Y	Y
Wikipedia (Wikimedia Foundation)	2,000 (as of September 2021) (Wikimedia, 2021[103])	Online encyclopaedia	N	N

Annex B - Profiles of the Top-50 Services

1. Facebook[93]

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 probably the Service with the most detailed explanation of what it deems terrorist organisations and terrorist content. In the section of Facebook's Community Standards entitled 'Dangerous Individuals and Organisations (Facebook, n.d.[33]), Facebook states that organisations or individuals that proclaim a violent mission or are engaged in violence cannot have a presence on Facebook. Facebook assesses these entities based on their behaviour both online and offline – most significantly, their ties to violence. The policy encompasses individuals, organisations and networks of people. They are divided into three tiers that indicate the level of content enforcement, with Tier 1 resulting in the most extensive enforcement because these entities have the most direct ties to offline harm. ***Tier 1: Terrorism, organised hate, large-scale criminal activity, mass and multiple murderers, and violating violent events*** Facebook does not allow individuals or organisations involved in organised crime, including those designated by the United States government as specially designated narcotics trafficking kingpins (SDNTKs); hate; or terrorism, including entities designated by the United States government as foreign terrorist organisations (FTOs) or specially designated global terrorists (SDGTs), to have a presence on the platform. Facebook also does not allow other people to represent these entities, nor leaders or prominent members of these organisations to have a presence on the platform, symbols that represent them to be used on the platform or content that praises them or their acts.

	In addition, Facebook removes any coordination of substantive support for these individuals and organisations. Facebook does not allow content that praises, substantively supports or represents events that Facebook designates as terrorist attacks, hate events, mass murders or attempted mass murders, serial murders, hate crimes or violating violent events. Praising, substantive support or representation of designated hateful ideologies are also prohibited. **Terrorist organisations and individuals** are defined as a 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. **Hate organisations** are defined as an association of three or more people that: • is organised under a name, sign or symbol; and • has an ideology, statements or physical actions that attack individuals based on characteristics, including race, religious affiliation, national origin, disability, ethnicity, gender, sex, sexual orientation or serious disease. **Criminal organisations** are defined as an association of three or more people that: • is united under a name, colour(s), hand gesture(s) or recognised indicia; and • has engaged in or threatens to engage in criminal activity such as homicide, drug trafficking or kidnapping.

	Mass and multiple murderers: • Facebook considers an event to be a mass murder or an attempted mass murder if it results in three or more casualties in one incident, defined as deaths or serious injuries. Any individual who has committed such an attack is considered to be a mass murderer or an attempted mass murderer. • Facebook considers any individual who has committed two or more murders over multiple incidents or locations a multiple murderer. **Hateful Ideologies** • Certain ideologies and beliefs are inherently tied to violence and attempts to organise people around calls for violence or exclusion of others based on their protected characteristics. In these cases, Facebook designates the ideology itself and remove content that supports this ideology from its platform. These ideologies include: o Nazism o White supremacy o White nationalism o White separatism • Facebook removes explicit praise, substantive support and representation of these ideologies, and remove individuals and organisations that ascribe to one or more of these hateful ideologies. ***Tier 2: Violent non-state actors*** Organisations and individuals designated by Facebook as violent non-state actors are not allowed to have a presence on Facebook, or have a presence maintained by others on their behalf. As these groups are actively engaged in violence, substantive support of these entities is similarly not allowed. Facebook also removes praise of violence carried out by these entities. **Violent non-state actors** are defined as any non-state actor that:

	• engages in purposive and planned acts of violence, primarily against a government military or other armed groups; and • that causes or attempts to o cause death to persons taking direct part in hostilities in an armed conflict, and/or o deprive communities of access to vital infrastructure and natural resources, and/or bring significant damage to property, linked to death, serious injury or serious harm to civilians ***Tier 3: Militarised social movements, violence-inducing conspiracy networks and hate-banned entities*** Pages, groups, events and profiles or other Facebook entities that are – or claim to be – maintained by, or on behalf of, militarised social movements and violence-inducing conspiracy networks, are prohibited. Admins of these Pages, groups and events will also be removed. Facebook does not allow representation of organisations and individuals designated by Facebook as hate-banned entities, e.g. **Militarised social movements (MSMs)**, which include: • **Militia groups**, defined as non-state actors that use weapons as a part of their training, communication or presence; and are structured or operate as unofficial military or security forces, and: o Coordinate in preparation for violence or civil war; or o Distribute information about the tactical use of weapons for combat; or o Coordinate militarised tactical coordination in a present or future armed civil conflict or civil war. • **Groups supporting violent acts amid protests**, defined as non-state actors that repeatedly: o Coordinate, promote, admit to or engage in:

	○ Acts of street violence against civilians or law enforcement; or ○ Arson, looting or other destruction of property; or ○ Threaten to violently disrupt an election process; or ○ Promote bringing weapons to a location when the stated intent is to intimidate people amid a protest. **Violence-inducing conspiracy networks (VICNs)**, defined as a non-state actor that: • Organises under a name, sign, mission statement or symbol; and • Promote theories that attribute violent or dehumanising behaviour to people or organisations that have been debunked by credible sources; and • Has inspired multiple incidents of real-world violence by adherents motivated by the desire to draw attention to or redress the supposed harms promoted by these debunked theories. **Hate-banned entities**, defined as entities that engage in repeated hateful conduct or rhetoric, but do not rise to the level of a Tier 1 entity because they have not engaged in or explicitly advocated for violence, or because they lack sufficient connections to previously designated organisations or figures. Also, in the section titled 'Violence and Incitement' of Facebook's Community Standards (Facebook, n.d.[104]), 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: ○ 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: ▪ 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: ○ 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 ● Statements of intent or advocacy, calls to action, or aspirational or conditional statements to bring weapons to locations, including but not limited to places of worship, educational facilities, polling places or locations to count votes or administer an election (or encouraging others to do the same). Lastly, Facebook now prohibits the posting of

	• Content that puts LGBTQI+ people at risk by revealing their sexual identity against their will or without permission. • Content that puts unveiled women at risk by revealing their images without veil against their will or without permission. • Violent threats against law enforcement officers. • Violent threats against people accused of a crime. Facebook removes this content when it has reason to believe that the content is intended to cause physical harm. • Misinformation and unverifiable rumours that contribute to the risk of imminent violence or physical harm. • Coded statements where the method of violence or harm is not clearly articulated, but the threat is veiled or implicit. Facebook looks at the below signals to determine whether there is a threat of harm in the content. o Shared in a retaliatory context (e.g. expressions of desire to do something harmful to others in response to a grievance or threat that may be real, perceived or anticipated) o References to historical or fictional incidents of violence (e.g. content that threatens others by referring to known historical incidents of violence that have been executed throughout history or in fictional settings) o Acts as a threatening call to action (e.g. content inviting or encouraging others to carry out harmful acts or to join in carrying out the harmful acts) o Indicates knowledge of or shares sensitive information that could expose others to harm (e.g. content that either makes note of or implies awareness of personal information that might make a threat of physical violence more credible. This includes implying knowledge of a person's residential address, their place of employment or education, daily commute routes or current location)

	○ Local context or subject matter expertise confirms that the statement in question could be threatening and/or could lead to imminent violence or physical harm. ○ The subject of the threat reports the content to us. • Threats against election officials • Implicit statements of intent or advocacy, calls to action, or aspirational or conditional statements to bring armaments to locations, including but not limited to places of worship, educational facilities, polling places or locations used to count votes or administer an election (or encouraging others to do the same). Facebook may also restrict calls to bring armaments to certain locations where there are temporarily signals of a heightened risk of violence or offline harm. This may be the case, for example, when there is a known protest and counter-protest planned or violence broke out at a protest in the same city within the last seven days.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://transparency.fb.com/en-gb/policies/community-standards/
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. For instance, someone who shares a link to a statement from a terrorist group with no context is immediately blocked from using Live for a set period of time.
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	If content goes against Facebook Community Standards (or Instagram Community Guidelines), Facebook will remove it. Facebook also notifies the user so they can understand why Facebook removed the content and how to avoid posting violating content in the future.
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. The assigned reviewer does not know that the post has been reviewed previously. 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. Currently, Facebook offers appeals for the vast majority of violation types on Facebook. Facebook does not offer appeals for violations with extreme safety concerns, such as child exploitation imagery (Facebook, 2021[105]). It is not stated whether this also applies to violations of the policy on dangerous individuals and organisations.
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 explains that technology helps them in three main areas: • **Proactive Detection:** Artificial intelligence (AI) has improved to the point that it can detect violations across a wide variety of areas without relying on users to report content to Facebook, often with greater accuracy than reports from users. This allows for the detection of harmful content, preventing it from being seen by hundreds or thousands of people. • **Automation:** AI has also helped scale the work of content reviewers. Facebook's AI systems automate decisions for certain areas where content is highly likely to be violating. This helps scale content decisions without sacrificing accuracy so that reviewers can focus on decisions where more expertise is needed to understand the context and nuances of a particular situation. Automation also makes it easier to take action on identical reports, so Facebook's teams do not

	have to spend time reviewing the same things multiple times. • **Prioritisation:** Instead of simply looking at reported content in chronological order, Facebook's AI prioritises the most critical content to be reviewed, whether it was reported to Facebook or detected by its proactive systems. This ranking system prioritises the content that is most harmful to users based on multiple factors such as virality, severity of harm and likelihood of violation. In an instance where Facebook's systems are near-certain that content is breaking its rules, it may remove it. Where there is less certainty, it will prioritise the content for teams to review (King, 2020[106]). Facebook states that its technology finds more than 90% of the content that it removes before anyone reports it (Facebook, 2021[107]). Facebook explains that it develops machine learning models to predict whether a piece of content is, for example, hate speech or violent and graphic content. Then, a separate system – its enforcement technology – determines whether to take an action, such as deleting, demoting or sending the content to a human review team for further review. A detailed explanation of how these systems are developed and implemented can be found at https://transparency.fb.com/en-gb/enforcement/detecting-violations/how-enforcement-technology-works/ and https://transparency.fb.com/en-gb/enforcement/detecting-violations/training-technology/ In particular, Facebook explains that when technology misses something or needs more input, then reviewers step in. As potential content violations get routed to review teams, each reviewer is assigned a queue of posts to individually evaluate. Sometimes, this review means simply looking at a post to determine whether it goes against Facebook's policies, such as an image containing adult nudity, in instances when technology did not detect it first. In other cases, context is key. For example, Facebook's technology might be unsure whether a post contains bullying, a policy area that requires extra context and nuance because it often reflects the nature of personal relationships. In this case, Facebook sends the post to review teams that have the right subject matter and language expertise for further review. If necessary, they can also escalate it to subject matter experts on the Global Operations or Content Policy teams (Facebook, 2021[108]). Explanations of the composition of Facebook's review teams, how they are trained, and the support tools at their disposal can

	be found at https://transparency.fb.com/en-gb/enforcement/detecting-violations/people-behind-our-review-teams/ , https://transparency.fb.com/en-gb/enforcement/detecting-violations/training-review-teams/ and https://transparency.fb.com/en-gb/enforcement/detecting-violations/making-the-right-calls/ Also, users have an option to flag content if they believe it violates Facebook's Community Standards. When a user reports content, it is routed through an automated system that determines how it should be reviewed. If this automated system determines that the content is clearly a violation, then it may be automatically removed. If the system is uncertain about whether the content is a violation, the content is routed to a human reviewer. 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	If content goes against the Facebook Community Standards (or Instagram Community Guidelines), Facebook will remove it. Facebook also notifies the user so they can understand why Facebook removed the content and how to avoid posting violating content in the future. Facebook uses a strike system to count violations and hold the user accountable for the content posted. Whether Facebook applies a strike depends on the severity of the content, the context in which it was shared and when it was posted. For most violations on Facebook, strikes will lead to the following restrictions: • **One strike:** Warning and no further restrictions. • **2 strikes:** One-day restriction from creating content, such as posting, commenting, using Facebook Live or creating a Page. • **3 strikes:** 3-day restriction from creating content. • **4 strikes:** 7-day restriction from creating content. • **5 or more strikes:** 30-day restriction from creating content. If content posted goes against Facebook's more severe policies, such as the policy on dangerous individuals and organisations, the violating user may receive additional, longer restrictions from certain features, on top of the standard restrictions above. For example, the user may be restricted

	from creating ads for set periods of time, starting on the first violation. Depending on which policy the content goes against, the user's previous history of violations and the number of strikes they have, their account may also be disabled (Facebook, 2021[109]). Facebook disables accounts as soon as it becomes aware of them, such as those of dangerous individuals and organisations (Facebook, 2021[110]).
7. Does the service issue transparency reports (TRs) specifically on content related to terrorism and/or violent extremism?	Yes (Facebook, 2017-2021[111]). 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'.
8. What information/fields of data are included in the TRs?	The latest report, issued in August 2021, includes the following five fields of information in both the 'Dangerous Organisations: Terrorism and Organised Hate' section and the 'Violence and Graphic Content' section. As for the former policy, metrics are broken down into Terrorism and Organised Hate. It must be noted that the report does not include data on other dangerous organisations prohibited from having a presence on Facebook and Instagram, including those engaging in mass or multiple murder, human trafficking or organized criminal activity: - *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. Facebook explains that views of violating content that contains terrorism are very infrequent, and it removes much of this content before people see it. As a result, many times there are not enough violating samples to precisely estimate prevalence. In Q2 2021, this was the case for violations of its policies on terrorism, suicide and self-injury and regulated goods on Facebook and Instagram. In these cases, Facebook can estimate an upper limit of how often someone would see content that violates these policies. In Q2 2021, the upper limit was 0.07% for violations of the policy for terrorism on Facebook. This means that out of every 10,000 views of content on Facebook, it is estimated that no more than 7 of those views contained content that violated the policy. Facebook also explains that currently it is unable to estimate prevalence for organised hate.

	- *Content actioned (How much content did Facebook take action on?)* Facebook indicates that a piece of content can be 'any number of things', including a post, photo, video or comment (Facebook, 2021[112]). 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. In the event that the content is escalated to law enforcement, Facebook does not additionally count that. 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). This includes content that Facebook actioned on after someone reported, and content Facebook found proactively. - *Proactive rate (Of the violating content actioned, how much did Facebook find before users reported it?)* This metric shows the percentage of content and accounts actioned for dangerous organisations and violence and graphic content that Facebook found and flagged before users reported it. The percentage of content flagged by users is also given. - *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. The metric is broken down into content restored after it is appealed, and restored after Facebook discovered issues itself (i.e. without appeal). Facebook also includes recent trends regarding content actioned for organised hate and terrorism. For example, its last transparency report notes that content actioned for terrorism decreased from 9 million pieces of content in Q1 2021 to 7.1 million in Q2 2021, and content actioned for organised hate decreased from 9.8 million pieces of content in Q1 2021 to 6.2 million in Q2 2021.
9. Methodologies for determining/ calculating/estimating the information/ data included in the TRs	- *Prevalence.* This metric assumes that the impact caused by violating content is proportional to the number of times that content is viewed. Prevalence of violating content is estimated using samples of content

	views from or across Facebook. It is calculated as the estimated number of views that showed violating content, divided by the estimated number of total content views on Facebook. 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. Facebook explains that some types of violations occur very infrequently. The likelihood that people view content that violate them is very low, and Facebook removes much of that content before people see it. As a result, many times Facebook does not find enough violating samples to precisely estimate prevalence. In these cases, Facebook can estimate an upper limit of how often someone would see content that violates these policies. For example, if the upper limit for terrorist propaganda was 0.04%, that means that out of every 10,000 views on Facebook in that time period, it is estimated that no more than four of those views contained content that violated Facebook's Terrorist Propaganda Policy. Facebook elaborates on the prevalence methodology in 'Prevalence' (Facebook, 2021[48]). - *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. In the event that the content is escalated to law enforcement, Facebook does not additionally count that. This metric includes both content Facebook actioned after someone reported it and content that Facebook found proactively. On Facebook, a post with no photo or video or a single photo or video counts as one piece of content. That means all of the following, if removed, would be counted as one piece of content actioned: a post with one photo, which is violating; a post with text, which is violating; and a post with text and one photo, one or both of which is violating. When a Facebook post has multiple photos or videos, we count each photo or video as a piece of content. For example, if Facebook removes two violating photos from a Facebook post with four photos, Facebook would count this as two pieces of content actioned: one for each photo removed. If Facebook removes the entire post, then Facebook counts the post as well. Thus, for example, if Facebook removes a Facebook post with four photos, Facebook would count this as five pieces of content actioned: one for each photo and one for the post. If Facebook only

	removes some of the attached photos and videos from a post, it only counts those pieces of content. At times, a piece of content will be found to violate multiple standards. For the purpose of measuring, Facebook attributes the action to only one primary violation. Typically, this will be the violation of the most severe standard. In other cases, the reviewer is asked to make a decision about the primary reason for violation. - *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. Facebook uses this metric as an indicator of how effectively it detects violations. - *Appealed Content.* This metric counts the number of pieces of content actioned for which people requested another review during the reporting period. Facebook reports the total number of pieces of content that had an appeal submitted in each quarter – for example, 1 January to 31 March. Thus, the numbers cannot be compared directly to content actioned or to content restored for the same quarter. Some restored content may have been appealed in the previous quarter, and some appealed content may be restored in the next quarter. It must be noted that Facebook's transparency report does not currently include any appeals metrics for accounts, Pages, groups and events that it took action on. - *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 reports content that it restored in response to appeals as well as content it restored that was not directly appealed. Facebook restores content without an appeal for a few reasons, including: • When Facebook made a mistake in removing multiple posts of the same content. In this case, Facebook only needs one person to appeal the decision to restore all of the posts. • When Facebook identifies an error in its review and restores the content before the person who posted it appeals.

	• When Facebook removes posts containing links that it identifies as malicious, and then learns that the link is not harmful anymore. In this case, Facebook can restore the posts.
10. Frequency/timing with which TRs are issued	As from August 2020, Facebook publishes its transparency reports on a quarterly basis. Its last report covers Q2 2021. Currently, there is available data from Q4 2017 to Q2 2021.
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 organisations 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. This includes, for example, videos that promote or glorify Nazi ideology, which is inherently discriminatory. YouTube also announced that it will remove content denying that well-documented violent events took place (Google/Youtube, 2019[113]). 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	If content violates any of YouTube's content policies, YouTube removes the content.

other enforcement decisions and appeal processes against them?	
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. 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, 2021[114]). 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, 2021[115]) 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. Also, when a video is removed, the user who posted it is given the opportunity to appeal. If the user chooses to submit an appeal, it goes to human review, and the decision is either upheld or reversed.

5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	YouTube relies on a combination of machine learning and people to detect problematic content at scale. Since machine learning is well suited to detect patterns, it allows for the detection of content that is similar to other that has already been remove, even before it is viewed. Also, YouTube recognises that the best way to quickly remove content is to anticipate problems before they emerge. Thus, its Intelligence Desk monitors the news, social media and user reports to detect new trends surrounding inappropriate content, and works to make sure that YouTube's teams are prepared to address them before they can become a larger issue (Google/YouTube, 2021[116]). YouTube also relies on the YouTube community to flag inappropriate content. In particular, users can use the flagging feature to this end. In addition, YouTube developed the YouTube Trusted Flagger Programme, which provides tools for individuals, government agencies and NGOs that are particularly effective at notifying YouTube of content that violates its Community Guidelines. Content flagged by Trusted Flaggers is not automatically removed or subject to any differential policy treatment — the same standards apply for flags received from other users. However, because of their high degree of accuracy, flags from Trusted Flaggers are prioritized for review by YouTube's teams (Google/YouTube, 2020[117]). 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 (Google/YouTube, 2020[117]). If content is not automatically removed by its machine learning systems, YouTube takes action on flagged videos after review by trained human reviewers. They assess whether the content does indeed violate YouTube's policies, and protect content that has an educational, documentary, scientific or artistic purpose. Reviewers' inputs are then used to train and improve the accuracy of YouTube's systems on a much larger scale (Google/YouTube, 2021[118]). With respect to the automated systems that detect extremist content, YouTube's teams have manually reviewed over two million videos to provide large volumes of training examples, which help improve the machine-learning flagging technology (Google/YouTube, 2021[119]). YouTube invests in a network of over 200 academics, government partners and NGOs which bring expertise to the platform's enforcement systems, including through YouTube's Trusted Flagger programme (Google/YouTube, 2021[119]). 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[120]), the Institute for Strategic Dialogue (ISDGlobal, n.d.[121]), the Wahid Institute in Indonesia and government agencies focused on counterterrorism. YouTube is a founding member of GIFCT and participates in the 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, they receive 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 • Show a trailer during Premieres • Send viewers from a live stream to a Premiere or send viewers from a Premiere to a live stream 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, 2021[122]). 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, 2021[123]). 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.[124]). YouTube issues transparency reports on the enforcement of its Community Guidelines. One section of these reports is about 'Violent Extremism' (Google/YouTube, 2021[119]). The last transparency report 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 transparency report also specifies that YouTube strictly prohibits content that promotes terrorism, such as content that glorifies terrorist acts or incites violence. In addition, the transparency report 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, 2010-2021[125]); • 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 removed videos that were 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) and by type of flagger (user, individual trusted flagger, NGO or government agency). • 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, 2021[119]); • the number of videos removed, by country/region • 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). • The 'violative view rate' (VVR), i.e. an estimate of the proportion of video views that violate YouTube's community guidelines in a given quarter (excluding spam). 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	Regarding the violative view rate, YouTube states that it first takes a sample of all videos that have been viewed on YouTube. The videos in that sample are then sent for review, and its teams determine whether each video does or does not violate its Community Guidelines. YouTube then use the aggregate results to estimate the proportion of views on YouTube that violate its Community Guidelines. The VVR metric is reported with a 95% confidence interval. This means that if measurement were performed many times for the same time period, YouTube

	would expect the true metric to lie within the interval 95% of the time (Google/YouTube, 2021[126]).
10. Frequency/timing with which TRs are issued	On a quarterly basis. Last TR covers Q2 2021.
11. Has this service been used to post TVEC?	Yes. See above sections 7-8.

3. Zoom

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 the section 'Terrorism and Violent Extremism Policy' of Zoom's Community Standards, Zoom states that terrorist or violent extremist groups on Zoom, or those who affiliate with them or promote their activities, are not allowed on Zoom.
	• Zoom defines terrorist organisations are those groups subject to national and international terrorism designations.
	• Zoom defines violent extremist groups are those groups 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.
	• Zoom will examine a group's activities both on and off Zoom to determine whether they engage in and/or promote violence against civilians to advance a political, religious and/or social cause.
	• Some specific examples of prohibited conduct under this policy are:
	  ○ engaging in or promoting acts on behalf of a terrorist organisation or violent extremist group;
	  ○ recruiting for a terrorist organization 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;

	○ using the insignia or symbols of terrorist organisations or violent extremist groups to promote them. In addition, in its 'Violent Threats Policy', Zoom states that violent threats or the glorification of violence on Zoom is prohibited. • Zoom believes that violent threats include statements of an intent to kill or inflict serious physical harm on a specific person or group of people. Stating an intent includes statements like "I will", "I'm going to", or "I plan to", as well as conditional statements like "If you do X, I will." Some examples of violent threats include: ○ threatening to kill someone; ○ threatening to sexually assault someone; ○ threatening to seriously hurt someone and/or commit a violent act that could lead to someone's death or serious physical injury; ○ asking for or offering a financial reward in exchange for inflicting violence on a specific person or group of people. Lastly, in its Hateful Conduct Policy, Zoom states that users cannot promote violence against, threaten, or harass other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. Users may not use their username, display name or profile information to abuse or threaten anyone. Moreover, there is no place on Zoom for organisations that promote violence against, threaten, or harass other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. • Zoom believes that hateful conduct is conduct that promotes violence against or directly attacks or threatens other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. • Zoom believes that hateful imagery includes logos, symbols, or images whose purpose is to promote hostility and malice against others based on their race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. Some examples of hateful imagery include:

	○ symbols historically associated with hate groups (e.g., the Nazi swastika); ○ images depicting others as less than human, or altered to include hateful symbols (e.g., altering images of individuals to include animalistic features); ○ images altered to include hateful symbols or references to a mass murder that targeted a protected category (e.g., manipulating images of individuals to include yellow Star of David badges, in reference to the Holocaust). • Violent threats include declarative statements of intent to inflict injuries that would result in death or serious and lasting bodily harm.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Zoom's Community Standards are available at https://explore.zoom.us/en/community-standards/
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?	Zoom states that it takes seriously its duty to safeguard the free and open exchange of thoughts and ideas on Zoom. Thus, it expects that all of its users observe the standards of behaviour that are included in its Community Standards.
4.1 Notifications of removals or other enforcement decisions	Zoom notifies account owners when actions are taken on their account. Violations of Community Standards that results in an account suspension or ban are informed the next time the relevant user attempts to use the platform.
4.2 Appeal processes against removals or other enforcement decisions	If the user believes Zoom's finding is wrong, they can submit an appeal. Zoom's appeal process can be found at https://zoom.us/appeals
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers,	\Zoom relies on reports to learn of alleged violations of its Community Standards.

hash-sharing/URL sharing database)	When a user makes a report about a violation of Zoom's Community Standards or Terms of Service, Zoom's Trust and Safety team will investigate and, if warranted, take action as quickly as possible. Zoom's tiered review process starts with a team of analysts who review different kinds of reports and flags in the first instance. Reports are first divided into queues by issue or reporter type. Team members rotate among the different queues so that everybody gain broad experience. As reports are resolved, the information about report type and resolution feeds into a dashboard. The dashboard gives Zoom meaningful data to spot trends, test abuse-prevention tools, or see spikes in demand so Zoom can refine our processes over time. Analysts escalate difficult or ambiguous cases to higher tiers. The highest tier is Zoom's Appeals Panel. Appeals Panelists serve for one-year terms and come from a diversity of backgrounds, experience levels, tenures, and departments at Zoom (Zoom, 2021[127]). Further details on this tiered process can be found at https://explore.zoom.us/docs/en-us/content-moderation-process.html?_ga=2.20044602.38595736.1624527871-1107759908.1602261224 Zoom uses automated tools to scan content such as virtual backgrounds, profile, images, and files uploaded or exchanged through chat for various categories of violations, including child sexual abuse material (CSAM), spam, violent extremism, and hateful conduct, among others (Zoom, 2021[128]). Zoom is a member of the GIFCT, but does not participate in the GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Accounts that violate Zoom's Community Standards may receive a strike, be suspended or permanently blocked, depending on the severity of the offence and prior conduct of the relevant user. Users who violate the Terrorism and Violent Extremism Policy and the Hateful Conduct Policy are permanently blocked.
7. Does the service issue transparency reports (TRs) on TVEC?	Yes (Zoom, 2021[127]).
8. What information/fields of data are included in the TRs?	Zoom's transparency reports include the following metrics: - Number and percentage of resolved reports by issue type (which includes terrorist or violent extremist groups) - Number and percentage of resolved reports by action taken - which may be user suspension, strike issued, onZoom/ZoomEvents host suspended, Event suspended,

	duplicate or dismissed. These actions have the following meanings: Dismissed: No action was taken. Duplicate: Two or more reports about the same issue from the same reporter. Event(s) Suspended: Zoom ended or prevented a particular event from taking place. OnZoom/Zoom Events Host(s) Suspended: Zoom blocked one or more hosts of OnZoom or Zoom Events. Strike Issued: The user received a strike. Strikes expire after 180 days and do not affect the user's ability to use the platform unless they accumulate. Depending on the reason for the strike, either one or two additional strikes within the same 180-day period will result in a suspension against the user. User(s) Suspended: The user was deactivated and/or blocked. They are prohibited from using Zoom unless they successfully appeal the decision.
9. Methodologies for determining/ calculating/estimating the information/ data included in the TRs	The data in Zoom's transparency report covers reports that it processed in a particular month, as opposed to reports that it received in a particular month.
10. Frequency/timing with which TRs are issued	On a monthly basis.
11. Has this service been used to post TVEC?	Yes. See Section 8 above.

4. 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, such as promoting violent crimes, endangering or exploiting children or others, or coordinating harm' (WhatsApp, 2021[129]).
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.whatsapp.com/legal/terms-of-service/?lang=en
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.
4.1 Notifications of removals or other enforcement decisions	If a number is banned, the user receives a notification.
4.2 Appeal processes against removals or other enforcement decisions	If a user believes that his or her account was terminated or suspended by mistake, the user can contact WhatsApp via email.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	WhatsApp states that it develops automated systems 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. 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 issuing TRs 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[130]).

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. 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.[131]) 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	Not applicable.

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?	Possibly. A security manual issued by ISIS recommended use of iMessage to protect supporters' identities, (Zetter, 2015[132]) but there is no evidence that ISIS supporters have actually used it (Dilger, 2015[133]). 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[134]).

6. Instagram

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	Facebook and Instagram share content policies. Facebook notes that if content is considered to be in violation of such policies on Facebook, it is also considered violating on Instagram. Therefore, Instagram follows the definitions set forth in Facebook's profile (see Section 1 of Facebook's profile). Because Facebook's Community Standards are more comprehensive than Instagram's Community Guidelines, they are the point of reference, even when considering Instagram violations. Instagram's Community Guidelines provide that Instagram is not a place to support or praise terrorism, organised 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. Also, serious threats of harm to public and personal safety are prohibited, as well as the sharing of graphic images to glorify violence.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Instagram's Community Guidelines are available at https://www.facebook.com/help/instagram/477434105621119/ 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	Instagram removes content from the platform when content violates its Community Guidelines.

(removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	
4.1 Notifications of removals or other enforcement decisions	If content goes against Instagram Community Guidelines, Instagram will remove it. Instagram also notifies the user so they can understand why Instagram removed the content and how to avoid posting violating content in the future.
4.2 Appeal processes against removals or other enforcement decisions	Instagram has the same appeals process as Facebook. See section 4.2 of Facebook's profile.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Instagram uses the same methods as Facebook to identify and remove objectionable content, including TVEC. See Section 5 of Facebook's profile. Instagram is a member of the GIFCT, and participates in the GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of ToS or Community Guidelines/Standards	If content goes against Instagram Community Guidelines, Instagram will remove it. Instagram also notifies the user so they can understand why Instagram removed the content and how to avoid posting violating content in the future. Depending on which policy the content goes against, the user's previous history of violations and the number of strikes they have, their account may also be restricted or disabled (Facebook, 2021[109]). Instagram follows the same approach as Facebook regarding sanctions (Facebook, 2021[109]) (see Section 6 of Facebook's profile).
7. Does the service issue transparency reports (TRs) on TVEC?	Yes. They are issued jointly with Facebook's (Facebook, 2017-2021[111]). The section of Instagram's report relevant to TVEC are 'Terrorism and Organised Hate' and 'Violence and Graphic Content'.
8. What information/fields of data are included in the TRs?	The latest report, issued in August 2021, includes the following five fields of information in both the 'Dangerous Organisations: Terrorism and Organised Hate' section and the 'Violence and Graphic Content' section. As for the former policy, metrics are broken down into Terrorism and Organised Hate. It must be noted that the report does not include data on other dangerous organisations prohibited from having a presence on Facebook and Instagram, including those engaging in mass or multiple murder, human trafficking or organized criminal activity: - *Prevalence (How prevalent were terrorism and violence and graphic content violations on Instagram?)* The prevalence

	metric is the percentage of views that included terrorism and violence and graphic content violations. Instagram explains that views of violating content that contains terrorism are very infrequent, and it removes much of this content before people see it. As a result, many times there are not enough violating samples to precisely estimate prevalence. In Q2 2021, this was the case for violations of its policies on terrorism, suicide and self-injury and regulated goods on Facebook and Instagram. In these cases, Instagram can estimate an upper limit of how often someone would see content that violates these policies. In Q2 2021, the upper limit was 0.05% for violations of the policy for terrorism on Facebook. This means that out of every 10,000 views of content on Instagram, it is estimated that no more than 5 of those views contained content that violated the policy. Instagram also explains that currently it is unable to estimate prevalence for organised hate. - *Content actioned (How much content did Instagram take action on?)* Instagram indicates that a piece of content can be 'any number of things', including a post, photo, video or comment (Facebook, 2021[112]). 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. In the event that the content is escalated to law enforcement, Instagram does not additionally count that. 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 guidelines (in this case the terrorism and violence and graphic content policies). This includes content that Instagram actioned on after someone reported, and content Instagram found proactively. - *Proactive rate (Of the violating content actioned, how much did Instagram find before users reported it?)* This metric shows the percentage of content and accounts actioned for dangerous organisations and violence and graphic content that Instagram found and flagged before users reported it. The percentage of content flagged by users is also given. - *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. The metric is broken down into

	content restored after it is appealed, and restored after Instagram discovered issues itself (i.e. without appeal). Instagram also includes recent trends regarding content actioned for organised hate and terrorism. For example, its last transparency report notes that content actioned for terrorism decreased from 429,000 pieces of content in Q1 2021 to 336,900 in Q2 2021, and content actioned for organised hate increased from 324,600 pieces of content in Q1 2021 to 367,300 in Q2 2021.
9. Methodologies for determining/ calculating/estimating the information/ data included in the TRs	- *Prevalence.* This metric assumes that the impact caused by violating content is proportional to the number of times that content is viewed. Prevalence of violating content is estimated using samples of content views from or across Instagram. It is calculated as the estimated number of views that showed violating content, divided by the estimated number of total content views on Instagram. 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. Instagram explains that some types of violations occur very infrequently. The likelihood that people view content that violate them is very low, and Instagram removes much of that content before people see it. As a result, many times Instagram does not find enough violating samples to precisely estimate prevalence. In these cases, Instagram can estimate an upper limit of how often someone would see content that violates these policies. For example, if the upper limit for terrorist propaganda was 0.04%, that means that out of every 10,000 views on Instagram in that time period, it is estimated that no more than four of those views contained content that violated Instagram's Terrorist Propaganda Policy. Instagram elaborates on the prevalence methodology in 'Prevalence' (Facebook, 2021[48]). - *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. In the event that the content is escalated to law enforcement, Instagram does not additionally count that. This metric includes both content Instagram actioned after someone reported it and content that Instagram found proactively. On Instagram, when a post contains violating content, the whole post is removed, and Instagram counts this as one piece of content actioned, regardless of how many photos or videos there are in the post. At times, a piece of content will be found to violate multiple standards. For the purpose of measuring, Instagram attributes the action to only one primary violation. Typically, this will be the violation of the most

	severe standard. In other cases, the reviewer is asked to make a decision about the primary reason for violation. - *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. Instagram uses this metric as an indicator of how effectively it detects violations. - *Appealed Content.* This metric counts the number of pieces of content actioned for which people requested another review during the reporting period. Instagram reports the total number of pieces of content that had an appeal submitted in each quarter – for example, 1 January to 31 March. Thus, the numbers cannot be compared directly to content actioned or to content restored for the same quarter. Some restored content may have been appealed in the previous quarter, and some appealed content may be restored in the next quarter. It must be noted that Instagram's transparency report does not currently include any appeals metrics for accounts, Pages, groups and events that it took action on. - *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 reports content that it restored in response to appeals as well as content it restored that was not directly appealed. Instagram restores content without an appeal for a few reasons, including: • When Instagram made a mistake in removing multiple posts of the same content. In this case, Instagram only needs one person to appeal the decision to restore all of the posts. • When Instagram identifies an error in its review and restores the content before the person who posted it appeals. • When Instagram removes posts containing links that it identifies as malicious, and then learns that the link is not harmful anymore. In this case, Instagram can restore the posts.
10. Frequency/timing with which TRs are issued	Instagram transparency reports 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[135]) (Hymas, 2019[136]) (Cox, 2019[137]).

7. 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. Facebook Messenger does not have specific ToS or Community Standards. However, as Facebook scans Facebook Messenger conversations to detect violations to its Community Standards, (Frier, 2018[138]) these Standards, which feature a well-developed description of terrorism and related concepts, apply to Facebook Messenger. See Section 1 of Facebook's profile.
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?	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 Facebook's profile.
4.1 Notifications of removals or other enforcement decisions	See Section 4.1 of Facebook's profile.
4.2 Appeal processes against removals or other enforcement decisions	See Section 4.2 of Facebook's profile.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	See Section 5 of Facebook's profile.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	See Section 6 of Facebook's profile.
7. Does the service issue transparency reports (TRs) on TVEC	See Section 7 of Facebook's profile.

8. What information/fields of data are included in the TRs?	See Section 8 of Facebook's profile.
9. Methodologies for determining/ calculating/estimating the information/ data included in the TRs	See Section 9 of Facebook's profile.
10. Frequency/timing with which TRs are issued	See Section 10 of Facebook's profile.
11. Has this service been used to post TVEC?	Yes. See above sections 7-8 of Facebook's profile.

8. 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 violent, criminal, illegal, or inappropriate content, as well as any content or behaviour that breaches any applicable laws or regulations. Such content or behaviour may include the following: • Threats to others, including statements of intent regarding committing violence (including murder or offering services for hire to kill others) or other criminal actions (e.g., kidnapping) • Instructions on how to make weapons or explosives • Misinformation that contributes to imminent violence or physical harm • Any organisation that promotes or is in the business/has the aim of promoting any illegal activities • Promoting or publicising violent crime, theft, and/or fraud • Facilitating or coordinating future criminal activity. WeChat also bans any organisations or persons who are involved in any of the above, including any related coordination or promotion. Criminal and/or illegal activities may include: • Terrorist activity, organised hate, kidnapping, human trafficking, or organised criminal activity

	• Violent acts – e.g., murder, harm against people or animals (excluding legal activities such as boxing, hunting or food preparation). • Offering of illegal goods or services – e.g. services for hire to kill others. WeChat also prohibits 'objectionable content and behaviour', which is defined as any content or behaviour that is reasonably likely to cause upset and/or distress, either to the subject and/or to the public. This may include: • Hate speech – e.g., a direct attack based on race, ethnicity, national origin, religion, sexual orientation, physical or mental disabilities, or other forms of 'dehumanising; speech or imagery, and • Graphic content of violence – including against both human beings and animals (and whether alive or dead).
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
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 seeks to ensure that any content which may (in its opinion) constitute a genuine risk of harm or direct threat to public safety is removed as soon as practicable, as well as any content that breaches any applicable laws or regulations.
4.1 Notifications of removals or other enforcement decisions	When content is removed, users are notified on the WeChat app.
4.2 Appeal processes against removals or other enforcement decisions	WeChat's content moderation decisions can be appealed (Ranking Digital Rights - Tencent Holdings Limited, 2021[139])
5. Means of identifying TVEC (for example, monitoring algorithms, user generated,	WeChat provides no information in this regard.

human (staff) reviewers, hash-sharing/URL sharing database)	It has been reported that Chinese online firms, including WeChat, have a team of moderators policing problematic content[94]. 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[140]). Also, research has shown that WeChat uses algorithmic technology (Knockel et al., 2018[53]), keyword filtering and URL blocking (Ruan, 2016[58]) 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, 2016[58]), 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[141]) WeChat is not a member of the GIFCT, and does not participate in the GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Where a user has breahed WeChat's Acceptable Use Policy, WeChat may, at its discretion, do any or all of the following: • Issue a warning regarding the user's behaviour. • Refrain from displaying or remove the relevant content relating to such breach (or reasonably suspected breach). • Display a notice to recipients of the relevant content to take precaution due to a suspected or confirmed breach of the policy. • Restrict the user from using certain account functions or suspend or terminate their account. • Where WeChat reasonably believes that the user has committed a crime or is otherwise required to do so under applicable laws, notify and cooperate with appropriate governmental and/or law enforcement authorities in the relevant jurisdiction.
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/	Not applicable.

estimating 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?	Yes. The Christchurch shooting was posted on WeChat (Kenny, 2019[142]). In addition, WeChat has been used to disseminate anti-Muslim propaganda (Huang, 2018[143]).

9. 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 Acceptable Use Policy states the following: Content that is related to terrorism, including the planning of a terrorist attack, promoting terrorist groups, is strictly prohibited. We may remove such content, disable accounts and work with law enforcement agencies (as necessary under applicable law) when we believe that there is a genuine risk of physical harm or a direct threat to public safety under such circumstances.
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/pt-pt/terms/viber-acceptable-use-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 if a user's Public Account is approved by Viber, the user automatically becomes a Public Account Administrator and Public Chat Administrator. Also, 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. 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	According to Viber's Acceptable User Policy: "We will make best efforts to notify the parties of our decision, however if we were not able to do so, you may appeal or contact our support."
4.2 Appeal processes against removals or other enforcement decisions	According to Viber's Acceptable User Policy: "In the event that we choose to take action against any particular user with respect to any content that he or she has posted or we decide to remove or refuse to distribute such content, you may appeal or contest our decision to remove content or disable, block or suspend your Account by contacting us through Viber Contact Us Form available here. Please include your reasoning as to why you feel our decision was incorrect. If we feel that our decision was in fact incorrect, we will notify you of such and rectify the situation by putting the content back, reactivating your account or the Services for you (as applicable) and removing any strikes or restrictions so that this will not be held against you in the future."
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	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. 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 channels are not end to end encrypted, and Viber can, upon reports, review them and if required remove them. Viber is not a member of the GIFCT, and does not participate in the GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	According to Viber's Acceptable User Policy: "We may remove the offending content, terminate or limit the visibility of your Account, or notify law enforcement. Viber may remove any Reported Content, at its sole discretion, if it finds it to be in breach of the AUP, Viber Terms or applicable law."
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[144]) a Nashir News Agency (the ISIS-linked media dissemination group) account on Viber (Katz, 2019[145]). Viber closed the account immediately after finding it.

10. 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 now feature detailed explanations of what content and organisations are banned from the platform: **Violent extremism** TikTok does not allow people to use the platform to threaten or incite violence, or to promote violent extremist individuals or organisations. **Threats and incitement to violence** It is understood as advocating for, directing, or encouraging other people to commit violence. TikTok does not allow threats of violence or incitement to violence on its platform that may result in serious physical harm. Thus, the following content is strictly prohibited: • Statements of intent to inflict physical injuries on an individual or a group • Statements or imagery that encourage others to commit or that advocate for physical violence • Conditional or aspirational statements that encourage other people to commit violence • Calls to bring weapons to a location with the intent to intimidate or threaten an individual or group with violence • Instructions on how to make or use weapons with an intent to incite violence

	Violent extremist organisations and individuals They are understood as individuals and organisations who promote or are engaged in violence. TikTok removes such individuals and organisations, including mass murderers, serial killers and rapists, hate groups, criminal organisations, terrorist organisations, and other non-state armed groups that target civilians. Terrorist organizations: terrorists and terrorist organisations are non-state actors that threaten violence, use violence, and/or commit serious crimes (such as crimes against humanity) against civilian populations in pursuit of political, religious, ethnic, or ideological objectives. Organised hate: it refers to those individuals and organisations who attack people based on protected characteristics, such as race, ethnicity, national origin, religion, caste, sexual orientation, sex, gender, gender identity, or immigration status. Attacks include actions that incite violence or hatred, dehumanise individuals or groups, or embrace a hateful ideology. Criminal organisations: these are transnational, national, or local groups that have engaged in serious crimes, including violent crimes (e.g., homicide, rape, robbery, assault), trafficking (e.g., human, organ, drug, weapons), kidnapping, financial crimes (e.g., extortion, blackmail, fraud, money laundering), or cybercrime. Thus, the following content is strictly prohibited: • Content that praises, promotes, glorifies, or supports violent extremist individuals and/or organisations • Content that encourages participation in, or intends to recruit individuals to, violent extremist organisations • Content with names, symbols, logos, flags, slogans, uniforms, gestures, salutes, illustrations, portraits, songs, music, lyrics, or other objects meant to represent violent extremist individuals and/or organisations **Hateful behaviour** TikTok does not permit content that contains hate speech or involves hateful behaviour. TikTok defines hate speech or behaviour as content that attacks, threatens, incites violence against, or otherwise dehumanises an individual or a group on the basis of protected attributes such as Race, Ethnicity, National origin, Religion, Caste, Sexual orientation, Sex, Gender, Gender identity, Serious disease,

	Disability and Immigration status. Thus, the following content is strictly prohibited:
	
	• Hateful content related to an individual or group, including:
	  o claiming that they are physically, mentally, or morally inferior
	  o calling for or justifying violence against them
	  o claiming that they are criminals
	  o referring to them as animals, inanimate objects, or other non-human entities
	  o promoting or justifying exclusion, segregation, or discrimination against them
	• Content that depicts harm inflicted upon an individual or a group on the basis of a protected attribute
	
	Slurs
	
	Slurs are defined as derogatory terms that are intended to disparage an ethnicity, race, or any other protected attributes listed above. To minimise the spread of egregiously offensive terms, TikTok removes all slurs from its platform, unless the terms are reappropriated, used self-referentially (e.g., in a song), or do not disparage.
	
	Hateful ideology
	
	Hateful ideologies are those that demonstrate clear hostility toward people because of their protected attributes. TikTok removes content that promotes hateful ideologies. Thus, the following content is strictly prohibited:
	
	• Content that praises, promotes, glorifies, or supports any hateful ideology (e.g., white supremacy, misogyny, anti-LGBTQ, antisemitism)
	• Content that contains names, symbols, logos, flags, slogans, uniforms, gestures, salutes, illustrations, portraits, songs, music, lyrics, or other objects related to a hateful ideology
	• Content that denies well-documented and violent events have taken place affecting groups with protected attributes

	• Claims of supremacy over a group of people with reference to other protected attributes • Conspiracy theories used to justify 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#39
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 states that it removes any content – including video, audio, livestream, images, comments, and text – that violates its Community Guidelines.. TikTok considers information available on other platforms and offline in its decisions to suspend or ban accounts. When warranted, TikTok reports the accounts to relevant legal authorities. TikTok's team of policy, operations, safety, and security experts work together to develop equitable policies that can be consistently enforced (TikTok, 2022[146]).
4.1 Notifications of removals or other enforcement decisions	Removals of any content, including video, audio, livestream, images, comments and text, are notified.
4.2 Appeal processes against removals or other enforcement decisions	TikTok offers creators the ability to appeal their video's removal. When it receives an appeal, it reviews the video a second time and reinstates it if it has been mistakenly removed.
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 human 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[147]). TikTok also has regional advisory councils beyond the US, including the Middle East, EU, Brazil and Asia Pacific[1]. TikTok is not a member of the GIFCT (although is currently applying for membership), and does not participate in GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	TikTok removes any content – including video, audio, livestream, images, comments, and text – that violates its Community Guidelines. Individuals are notified of our decisions and can appeal if they believe no violation has occurred. TikTok suspends or ban accounts and/or devices that are involved in severe or repeated violations (see https://newsroom.tiktok.com/en-us/advancing-our-approach-to-user-safety); TikTok considers information available on other platforms and offline in these decisions. When warranted, TikTok reports the accounts to relevant legal authorities. When TikTok finds an account that belongs to a violent and extremist organisation or individual, Tiktok closes it immediately.

[1] See https://newsroom.tiktok.com/en-gb/tiktok-european-safety-advisory-council ,

https://newsroom.tiktok.com/en-sg/tiktok-apac-safety-advisory-council ,

https://newsroom.tiktok.com/pt-br/tiktok-apresenta-seu-conselho-consultivo-de-seguranca-do-brasil and

https://newsroom.tiktok.com/ar-mena/tiktok-establishes-first-menat-safety-advisory-council-to-guide-safety-best-practice-and-policy

7. Does the service issue transparency reports (TRs) on TVEC?	Yes (TikTok, 2022[146]).
8. What information/fields of data are included in the TRs?	TikTok's last transparency report, which covers Q3 2021, includes the following metrics: - The number of videos removed globally for violating TikTok's Community Guidelines and/or ToS. The five countries with the largest volumes of removed videos are also reported; - The number and percentage of videos that were proactively caught and removed by TikTok's systems before a user reported them; - The number of videos reinstated after appeal; - The percentage of videos removed by removal reason (including violent extremism, hateful behaviour and violent and graphic content); - The proactive removal rate, removal rate before any views and removal within the 24 hours rate by removal reason (including violent extremism, hateful behaviour and violent and graphic content). Proactive removal means identifying and removing a violative video before it is reported to TikTok. Removal within 24 hours means removing the video within 24 hours of it being posted; and - The number of accounts removed for violating the Community Guidelines or Terms of Service. On violent extremism in particular, TikTok informs that in Q3 2021, of all videos removed, 0.89% violated this policy, compared to 1.1% in the Q2 2021. Of these videos, 92.08% were removed before they were reported, and 91.62% were removed within 24 hours of being posted. TikTok believes this increase in removals is due to its guidelines now describing in greater detail what is considered a violent threat and/or incitement to violence in terms of the content and behaviour it prohibits (TikTok, 2022[146]).
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	Since 2021 transparency reports will be released on a quarterly basis.

11. Has this service been used to post TVEC?	Yes, see Sections 7-8 above.

11. 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[95]
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?	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 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	Users can report violations through the reporting function. QQ also uses tools to proactively discover policy violations.

generated, human (staff) reviewers, hash-sharing/URL sharing database)	China's Cybersecurity law requires Internet-based companies to monitor user-generated content for information that is 'prohibited from being published or transmitted by laws or administrative regulations'. Companies are bound to invest in staff and filtering technologies to moderate content and remain compliance with government regulations (Ruan, 2019[148]). 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.

12. 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 https://terms.alicdn.com/legal-agreement/terms/suit_bu1_unification/suit_bu1_unification202005142208_14749.html?spm=a2hbt.13141534.app.5~5!5~5~5~DL!2~5~A
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?	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. China's Cybersecurity law requires Internet-based companies to monitor user-generated content for information that is 'prohibited from being published or transmitted by laws or administrative regulations'. Companies are bound to invest in staff and filtering technologies to moderate content and remain compliance with government regulations (Ruan, 2019[148]). 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.

13. 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 enforcement decisions and appeal processes against them?	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 that so far, this has never happened (Telegram, n.d.[149]).
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. Moderators review the reports and take action accordingly. 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.[150]). 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. However, on its ISIS Watch channel, Telegram discloses the number of ISIS terrorist bots and channels it bans every day, and the aggregate monthly number (e.g. 893 terrorist bots and channels banned on 12 October 2021, total this month: 13220).
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[151]) (Hayden, 2019[152]) (Bennett, 2019[151]) (Hayden, 2019[152]).

14. 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[96] 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 [97]
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. China's Cybersecurity law requires Internet-based companies to monitor user-generated content for information that is 'prohibited from being published or transmitted by laws or administrative regulations'. Companies are bound to invest in staff and filtering technologies to moderate content and remain compliance with government regulations (Ruan, 2019[148]). QQ International is not a member of the GIFCT, and does not participate in the 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 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.

15. 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. China's Cybersecurity law requires Internet-based companies to monitor user-generated content for information that is 'prohibited from being published or transmitted by laws or administrative regulations'. Companies are bound to invest in staff and filtering technologies to moderate content and remain compliance with government regulations (Ruan, 2019[148]). 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[142]).

16. 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, hate groups and hate speech', Snap states that • Terrorist organisations and hate groups are prohibited from using its platform; there is no tolerance for content that advocates or advances violent extremism or terrorism. • Hate speech or content that demeans, defames or promotes discrimination or violence on the basis of race, colour, caste, ethnicity, national origin, religion, sexual orientation, gender identity, disability or veteran status, immigration status, socio-economic status, age, weight or pregnancy status is prohibited. Also, under the heading 'Threats, violence & harm', Snap states that encouraging violence or dangerous behaviour is prohibited. 'Snaps' of gratuitous or graphic violence are not allowed.
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 has no live-streaming 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?	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 supports the Santa Clara Principles on Transparency and Accountability in Content Moderation (Santa Clara University's High Tech Law Institute, n.d.[52]), 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	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 content that violates Snapchat's policies (Snap Inc., n.d.[153]). In addition to in-app reporting, Snap also offers online reporting options through its support site. Furthermore, its teams work to improve capabilities for proactively detecting violating and illegal content, such as child abuse material, content that involves illegal drugs or weapons or threats of violence (Snap Inc., 2021[154]) Snap has a dedicated trust and safety team working on a 24/7 basis. The team reviews user reports to determine whether there is a violation of the Community Guidelines and whether any action needs to be taken. Snapchat is not a member of the GIFCT, but does participate in the GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	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?	Yes. Snap's last transparency report (Snap Inc., 2021[154]) features for the first time information on removals of TVEC.
8. What information/fields of data are included in the TRs?	Snap's transparency report includes:

	- Country-by-country information (including TVEC) - Total number of content reports; - The total amount of pieces of content that were enforced against for violation of Snap's Community Guidelines during the reporting period; - Total number of unique accounts enforced; - The number of content reports, content enforced, turnaround time and unique accounts enforced are broken down by type of policy violation (which include threatening / violence / harm and hate speech); - The violative view rate (VVR), i.e. the proportion of all Snaps (or views) that contained content that violated Snap's Community Guidelines during the reporting period. During the last reporting period, the VVR was 0.08 percent, which means that out of every 10,000 views of content on Snapchat, 8 contained content that violated its guidelines; - The number of account removals for violations of Snap's prohibition of terrorism, hate speech and extremist content. Snap indicates that it both its product architecture and the design of its Group Chat functionality limits the spread of TVEC and opportunities to organise. Snap offers Group Chats, but they are limited in size to several dozen members, are not recommended by algorithms, and are not discoverable on the platform if a user is not a member of that Group. Snap monitors developments in this area and mitigates any potential vectors for abuse on its platform (Snap Inc., 2021[154]).
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Enforcement actions could include removing the offending content; terminating or limiting the visibility of the account in question; and referring the content to law enforcement. Turnaround time reflects the median time in hours to action on a user report.
10. Frequency/timing with which TRs are issued	On a semi-annual basis.
11. Has this service been used to post TVEC?	Yes. See section 8 above.

17. 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. China's Cybersecurity law requires Internet-based companies to monitor user-generated content for information that is 'prohibited from being published or transmitted by laws or administrative regulations'. Companies are bound to invest in staff and filtering technologies to moderate content and remain compliance with government regulations (Ruan, 2019[148]). 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.

18. 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. China's Cybersecurity law requires Internet-based companies to monitor user-generated content for information that is 'prohibited from being published or transmitted by laws or administrative regulations'. Companies are bound to invest in staff and filtering technologies to moderate content and remain compliance with government regulations (Ruan, 2019[148]). 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.

19. 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, under the 'Violent actors' heading of Pinterest's Community Guidelines, Pinterest states that its platform is not a place for violent content, groups or individuals. Pinterest limits the distribution of or remove content and accounts that encourage, praise, promote or provide aid to dangerous actors or groups and their activities. This includes: • Extremists • Terrorist organisations • Gangs and other criminal organisations

	These terms are not defined. Also, under the 'Hateful activities' heading of Pinterest's Community Guidelines, Pinterest states that it removes hateful content or accounts of people and groups that promote hateful activities, such as: • 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 states that its platform is not a place for antagonistic, explicit, false or misleading, harmful, hateful, or violent content or behaviour. Thus, it may remove, limit or block the distribution of such content and the accounts, individuals, groups and domains that create or spread it based on how much harm it poses.
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.[155]).
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. Pinterest is a member of the GIFCT, and participates in the 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?	Yes. In its last transparency report (which covers Oct-Dec 2020), Pinterest discloses its enforcement actions for violations of its Community Guidelines, broken down by type of policy violation (including 'violent actors' and 'hate activity).
8. What information/fields of data are included in the TRs?	The following metrics are included in Pinterest' last transparency report: - Reach of policy-violating Pins (content with a message): percentage of Pins seen by 0 people, <10 people, 10-100 people and 100+ people, by type of policy violation; - Number of actioned used reports, by type of policy violation; - Number of distinct images and Pins deactivations, by type of policy violation;

	- Number of board deactivations, by type of policy violation; - Number of account deactivations, by type of policy violation; - Number of account appeals and reinstatements, by type of policy violation.
9. Methodologies for determining/ calculating/estimating the information/data included in the TRs	Users can report any content they find objectionable by clicking on the three small dots on any Pin and hitting 'Report Pin'. Once it is confirmed that it is a policy violation and Pinterest takes action on the reported content, Pinterest considers the report an actioned user report. There are more than 300 billion Pins on Pinterest, and each of those Pins also has an image associated with it. Just because two Pins show the same image does not mean that Pinterest counts them as the same Pin within its systems, or even that the image came from the same source. Thus, if Pinterest determines that the image in one Pin is policy-violating, its tools need to be able to detect and act on matching images amongst the billions of other Pins on Pinterest. Accordingly, whilst Pinterest detects and deactivates a lot of Pins, those Pins comprise a much smaller number of distinct images. That is why Pinterest shares the number of deactivations for both distinct images and Pins. When people find Pins they like or want to come back to, they save them to boards that they have created. Over time, people have created more than 6 billion boards. Pinterest deactivates boards if a predetermined amount of content on that board has been identified as policy-violating. When a board is deactivated, all the Pins on that board are also deactivated.
10. Frequency/timing with which TRs are issued	On a semi-annual basis.
11. Has this service been used to post TVEC?	Unknown (there are no reported violations of the 'violent actors' policy).

20. 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: **Threatening violence** Content that encourages, glorifies, incites, or calls for violence

	or physical harm against an individual (including oneself), a group of people, or animals. **Hate** Vilifying, humiliating, harassing, promoting identity-based attacks against, promoting hatred against, or threatening violence against marginalised or vulnerable groups. **NetzDG violations** Behaviour or content that's in violation of the specific sections of the German Criminal Code identified in the Network Enforcement Act (NetzDG) (See Section 4 of the report above). In its transparency report, Reddit clarifies that the TVEC it removes relates to US-designated foreign terrorist organisation (Reddit, 2022[156]).
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 Reddit's Moderator Guidelines are available at https://www.redditinc.com/policies/moderator-guidelines 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 Additionally, 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.[157]) 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). Communities may appeal to be removed from the quarantine state.
4.1 Notifications of removals or other enforcement decisions	Individual content removals or account suspensions are notified via a private message. Subreddit removals are tombstoned with the removal reason so that visitors may see why and when the community was banned.
4.2 Appeal processes against removals or other enforcement decisions	Enforcement actions taken by Reddit administrators, whether against a community or an individual user, may be appealed. Reddit discloses detailed information about appeals statistics in its annual Transparency Report. Reddit's Moderator Guidelines also require that individual subreddits provide for appeal of volunteer moderator actions. They may manage these appeals mechanisms within their particular communities at their own discretion. As such, the appeals process will vary from subreddit to subreddit.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Reddit has a community-based, federal-style approach to content moderation that shares moderation responsibility between company employees and 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[158]), 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.[159]). In addition, specially trained Reddit employees are in charge of enforcing Reddit's Content Policy at the sitewide level. They especially focus on violations at scale (spam or other coordinated attacks) and complex situations that require access

	to backend data or tools, such as hash-matching technology. They also take action when violations demand a higher-level response than moderators are capable of, such as banning a user from the entire site, removing an entire subreddit, or appropriately addressing illegal material. 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 is not a member of the GIFCT, but does participate in the GIFCT's Hash Sharing Consortium, employing automated detection methods against this hash set.
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 does issue 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. In its last report (2021), Reddit specifically reported that out of the total amount of violent content removed (17,487 pieces of content), there were 97 pieces of US-designated foreign terrorist organisation content (Reddit, 2022[156]) In its 2021 report (Reddit, 2022[156]), Reddit explained that the vast majority of 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. 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; - the volume of content removed by moderators is broken down by removal proactively performed by a human moderator or by Automod; - the number and percentage of Content Policy violations removed by Reddit administrators is divided by categories of violations (Content manipulation, Harassment, Minor sexualization, Violent content, Hateful Content, Involuntary pornography, Prohibited goods, Personal identifiable information, Impersonation, or Other). It is also broken down by type of content (post & comment, private message, chat); - What percentage of removals per rule were surfaced via user reports vs automation; - the number of subreddits removed by categories of removal reason (Content manipulation, hateful content, personally identifiable information, prohibited goods, involuntary porn, minor sexualisation, violent content, harassment, ban evasion, trademark, copyright, unmoderated); - the number of accounts sanctioned (broken down by warnings, temporary bans and permanent bans) by Reddit administrators, by type of violations of the Content policy (Content manipulation, ban evasion, Harassment, Minor sexualization, Violent content, Hateful Content, Involuntary pornography, Prohibited goods, Personal identifiable information, copyright, trademark, and Impersonation); - the number of quarantined subreddits; - the number of reports Reddit received for potential policy violations, and the percentage of such reports that resulted in action taken by Reddit Administrators (“actionability”); and - total number of appeals received by Reddit, listed by rule and broken down into appeals granted and denied;

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[160]) This led to Reddit administrators banning the entire community in question from the site. See also Section 7 above.

21. Twitter

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 terrorist or violent extremist *content*, but there is a specific policy on 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. Also, Twitter considers as 'other violent organisations' those that meet the following criteria:

	• a collection of individuals with a shared purpose; and • have systematically targeted civilians with violence. 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 or indicate affiliation or support. Twitter updated its Hateful conduct policy in September 2020 to expand Twitter's enforcement approach towards content that incites fear/fearful stereotypes about protected categories, in part due to trends noticed about the targeting of certain protected categories in light of the COVID-19 pandemic. Hateful conduct also expanded to include a new dehumanization policy update to cover content that dehumanises on the basis of race, ethnicity, or national origin. This policy applies to: **Violent threats:** these are declarative statements of intent to inflict injuries that would result in serious and lasting bodily harm, where an individual could die or be significantly injured, e.g., "I will kill you." **Wishing, hoping or calling for serious harm on a person or group of people:** Content that wishes, hopes, promotes, incites, or expresses a desire for death, serious bodily harm, or serious disease against an entire protected category and/or individuals who may be members of that category is prohibited. This includes, but is not limited to: • Hoping that an entire protected category and/or individuals who may be members of that category

	dies as a result of a serious disease, e.g., "I hope all [nationality] get COVID and die." • Wishing for someone to fall victim to a serious accident, e.g., "I wish that you would get run over by a car next time you run your mouth." • Saying that a group of individuals deserve serious physical injury, e.g., "If this group of [slur] don't shut up, they deserve to be shot." • Encouraging others to commit violence against an individual or a group based on their perceived membership in a protected category, e.g., "I'm in the mood to punch a [racial slur], who's with me?" **References to mass murder, violent events, or specific means of violence where protected groups have been the primary targets or victims:** Twitter prohibits targeting individuals or groups with content that references forms of violence or violent events where a protected category was the primary target or victims, where the intent is to harass. This includes, but is not limited to media or text that refers to or depicts: • genocides, (e.g., the Holocaust); • lynchings. **Incitement against protected categories:** Twitter prohibits inciting behaviour that targets individuals or groups of people belonging to protected categories. This includes content intended: • to incite fear or spread fearful stereotypes about a protected category, including asserting that members of a protected category are more likely to take part in dangerous or illegal activities, e.g., "all [religious group] are terrorists." • to incite others to harass members of a protected category on or off platform, e.g., "I'm sick of these [religious group] thinking they are better than us, if any of you see someone wearing a [religious symbol of the religious group], grab it off them and post pics!" • to incite others to discriminate in the form of denial of support to the economic enterprise of an individual or group because of their perceived membership in a protected category, e.g., "If you go to a [religious group] store, you are supporting those [slur], let's stop

	giving our money to these [religious slur].” This may not include content intended as political in nature, such as political commentary or content relating to boycotts or protests. **Repeated and/or non-consensual slurs, epithets, racist and sexist tropes, or other content that degrades someone:** Twitter prohibits targeting others with repeated slurs, tropes or other content that intends to dehumanise, degrade or reinforce negative or harmful stereotypes about a protected category. This includes targeted misgendering or deadnaming of transgender individuals. Twitter also prohibits the dehumanisation of a group of people based on their religion, caste, age, disability, serious disease, national origin, race, or ethnicity. **Hateful imagery:** hateful imagery is considered to be 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. Some examples of hateful imagery include, but are not limited to: • symbols historically associated with hate groups, e.g., the Nazi swastika; • images depicting others as less than human, or altered to include hateful symbols, e.g., altering images of individuals to include animalistic features; or • images altered to include hateful symbols or references to a mass murder that targeted a protected category, e.g., manipulating images of individuals to include yellow Star of David badges, in reference to the Holocaust.
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, and https://help.twitter.com/en/rules-and-policies/hateful-conduct-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	Twitter has a range of enforcement options that it may exercise when a user violates the Twitter Rules (Twitter, n.d.[161]).

content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	a. *Tweet-level enforcement*: applies to content that violates Twitter's policies, but Twitter believes it is in the public interest that such content remains accessible (i.e. the public interest exception). 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 labelling a tweet that may contain disputed or misleading information, 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 a notice 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.[162]).
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 suspends 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 is member of the GIFCT and participates in the 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.

7. Does the service issue transparency reports (TRs) on TVEC?	Yes. Twitter's Transparency Reports (Twitter, 2012-2021[163]) 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; • 'Accounts reported': the number of unique accounts that were reported 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 and Hateful Conduct). In its last transparency report, Twitter introduced a metric called 'impressions', which capture the number of views a Tweet received prior to removal. Twitter reports that from 1 July 2020 through 31 December 2020, Twitter removed 3.8M Tweets that violated the Twitter Rules. Of the Tweets removed, 77% received fewer than 100 impressions, with an additional 17% receiving between 100 and 1000 impressions. Only 6% of removed Tweets had more than 1000 impressions. In total, impressions on violative Tweets accounted for less than 0.1% of all impressions for all Tweets during that time period. Specifically for Terrorism and Violent Extremism, Twitter reports the percentage of actioned accounts which were proactively identified and actioned (96% of the 58,750 unique accounts actioned under the policy during the last reporting period 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 35% 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. 'Impression' is defined as any time at least half of the area of a given Tweet is visible to a user for at least half a second (including while scrolling). This also includes views by logged-out users.
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.

22. 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, it 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	Tumblr notifies users when it finds there has been a violation of its policies, at its discretion.
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.

	Tumblr has observed that the recently introduced Election Integrity policy in Tumblr's Community Guidelines has helped address some far-right violence extremism on Tumblr. Tumblr is a member of the GIFCT, but does not participate in the GIFCT's Hash Sharing Consortium.
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[164]), 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[165]). In 2019 Tumblr was sold to Automattic. Several Tumblr Transparency Reports have been published ever since, but none of them contain any information on TVEC (Tumblr, 2013-2020[166])
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. Pages promoting Nazism, white supremacy, ethno-nationalism, and far-right terrorism have been found on Tumblr (Barnes, 2019[167]) (Fisher-Birch, 2018[168]). Tumblr has ever since strived to improve its content moderation efforts, joining Tech Against Terrorism and the GIFCT.

23. LinkedIn

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No definition of TVEC is provided. However, LinkedIn's Professional Community Policy explicitly ban TVEC and associated activities: **Do not post terrorist content or promote terrorism:** LinkedIn does not allow any terrorist organisations or violent extremist groups on its platform. Also, LinkedIn does not allow any individuals who affiliate with such organisations or groups to

	promote their activities. Content that depicts terrorist activity, that is intended to recruit for terrorist organisations, or threatens, promotes, or supports terrorism in any manner is not tolerated. **Do not be hateful:** LinkedIn does not allow content that attacks, denigrates, intimidates, dehumanises, incites or threatens hatred, violence, prejudicial or discriminatory action against individuals or groups because of their actual or perceived race, ethnicity, national origin, caste, gender, gender identity, sexual orientation, religious affiliation, or disability status. Hate groups are not permitted on LinkedIn. Use of racial, religious, or other slurs that incite or promote hatred, or any other content intended to create division, is prohibited. **Do not threaten, incite, or promote violence:** LinkedIn does not allow threatening or inciting violence of any kind. LinkedIn does not allow individuals or groups that engage in or promote violence, property damage, or organised criminal activity. LinkedIn cannot be used to express support for such individuals or groups or to otherwise glorify violence.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.linkedin.com/help/linkedin/answer/34593 (click "Learn more about being safe")
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 at: https://www.linkedin.com/help/linkedin/ask/lv-app LinkedIn has provided additional best practices and guidelines for live streaming, which are available at: 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 warning or suspending the author if the content is in violation of its ToS or policies.
4.1 Notifications of removals or other enforcement decisions	Content removals are notified.

4.2 Appeal processes against removals or other enforcement decisions	If an account has been restricted or content has been removed and the user believes the action was in error, the user 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. Whenever terrorist content on LinkedIn is brought to its attention via its online reporting tool, LinkrfIn removes such content. In addition, LinkedIn employs machine classifiers and processers to detect potential TVEC on its platform. LinkedIn is a member of the GIFCT and participates in the GIFCT's Hash Sharing Consortium.
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 a warning or 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 at https://about.linkedin.com/transparency/community-report#content-violations
8. What information/fields of data are included in the TRs?	LinkedIn's last transparency report, which covers the period January to June 2021, discloses the total number of pieces of content removed by type of policy violation, including the 'violent or graphic' category which encompasses 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. There is no specific information on removals of TVEC.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Broad explanations are provided in the Community Report available at https://about.linkedin.com/transparency/community-report

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[169]).

24. 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. China's Cybersecurity law requires Internet-based companies to monitor user-generated content for information that is 'prohibited from being published or transmitted by laws or administrative regulations'. Companies are bound to invest in

	staff and filtering technologies to moderate content and remain compliance with government regulations (Ruan, 2019[148]). 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.

25. 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. China's Cybersecurity law requires Internet-based companies to monitor user-generated content for information that is 'prohibited from being published or transmitted by laws or administrative regulations'. Companies are bound to invest in staff and filtering technologies to moderate content and remain compliance with government regulations (Ruan, 2019[148]). Baidu Tieba is not a member of the GIFCT, and does not participate in the 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.

26. 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. Also, Quora imposes the following prohibitions: **No glorifying or advocating violence** Quora does not allow content that glorifies violence. Images, videos, and descriptions of violence should not be added with the intent to traumatise others or make them uncomfortable. **No hate speech** Quora is a place for civil discourse and does not tolerate content that attacks or disparages an individual or group based on race, gender, religion, nationality, ethnicity, political group, sexual orientation or another similar characteristic. Any generalisations about these topics should be phrased as neutrally as possible.
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://help.quora.com/hc/en-us/articles/360000470706-What-is-Quora-s-Be-Nice-Be-Respectful-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?	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.[170]) 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. If an edit-block or banning decision is difficult and/or involves a person who is active on Quora, then the decision will be made collectively by the admins as a group with each admin having the opportunity to provide input. Quora is not a member of the GIFCT, and does not participate in the 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[171]).

27. Microsoft Teams

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No express definition of TVEC is provided. However, Microsoft's Services Agreement, which governs Teams, 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[172]) that, for the purposes of its services, terrorist content is material posted by or in support of organisations included on the Consolidated United Nations Security Council Sanctions List (United Nations Security Council, n.d.[173]) that depicts graphic violence, encourages violent action, endorses a terrorist organisation 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 organisations.
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[98].
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 Microsoft's Services Agreement, violation of said agreement may result in Microsoft stopping the provision of services to the infringer, or closing their Microsoft account. Microsoft may also block delivery of a communication (like email, file sharing or instant message) to or from the Services (which include Microsoft Teams), or it may remove or refuse to publish a user's content for any reason. When investigating alleged violations of the Services Agreement, Microsoft reserves the right to review user's 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 deploys a variety of scanning technology, artificial intelligence, external partnerships, and human moderation operations solutions to detect and investigate TVEC. Furthermore, users are able to report abusive comments or content (Microsoft, n.d.[174]) Moderators review the reports 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[172]). Microsoft is a founding member of the GIFCT and participates in the GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	See information in Section 4 above.
7. Does the service issue transparency reports (TRs) on TVEC	Yes. TVEC numbers for Microsoft Teams are included in Microsoft's Digital Safety Content Report (Microsoft, 2020-2021[175]). This report is inclusive of Microsoft consumer products and services including (but not limited to) OneDrive, Outlook, Skype, Bing and Xbox[99]. 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 TVEC actioned reported by users or third parties • % 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	On a semi-annual basis.
11. Has this service been used to post TVEC?	Unknown.

28. 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 distribute content that promotes bigotry, racism, misogyny, and religious or ethnic hatred. Violent content is also prohibited. Also, IMO's Community Guidelines has explicit prohibitions of Terrorism/violent extremism and Hateful Speech: **Terrorism/violent extremism:** The production and distribution of any media that promotes terrorism or violent extremism, including but not limited to terrorism or extremism tendencies, statements, photographs of terrorist leaders, media content related to hostage-taking by extremists, bloody violence content, etc, is prohibited. **Hateful speech:** Users may not attack anyone based on their race, ethnicity, national origin, gender, gender identity, sexual orientation, religious affiliation, disabilities, or diseases.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://imo.im/policies/terms_of_service , https://imo.im/policies/acceptable_use_policy.html and https://imo.im/policies/community_guidelines.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. to IMO indicates that, to avoid malicious acts such as posting harmful content that could impact its community, IMO deploys advanced artificial intelligence technology to detect that content on a 24/7 basis (IMO, n.d.[176]). Users can report content that violates IMO's community guidelines by clicking on the report button on the relevant features. IMO has a global team of reviewers working on a 24/7 basis. Reviewers assess the reports and remove content and accounts that do not meet IMO's guidelines. IMO is not a member of the GIFCT, and does not participate in the 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. With regards to terrorist/extremist content and hateful speech in particular, IMO states that said content is removed and accounts may be temporarily or permanently suspended if cases are confirmed.
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. 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 may not post content which is racist, sexist, ageist or discriminatory in any other way.
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	When Ask.fm reviews a profile, it removes violating content and sends a warning message if the user's profile has a higher rate of violations during the last visits on Ask.fm.
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. Ask.fm uses 'pre-moderation tools' to detect harmful content automatically. This involves the use of a pattern system that distinguishes web-links, words, and expressions as hurtful or suspicious. In 2020, Ask.fm enriched its pattern list with 33,734 patterns. When a new threat on the platform is defined, Ask.fm adds new variation patterns in the system to prevent it from appearing again. Ask.fm also creates patterns for exceptional events happening in the world to detect text content that can constitute a danger to its users (Ask.fm, 2021[44]). Ask.fm is not a member of the GIFCT, but does participate in the GIFCT's Hash Sharing Consortium.
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?	Perhaps. Ask.fm participates in the GIFCT's Hash Sharing Consortium, and recently published a transparency report covering 2020 (Ask.fm, 2021[44]). The report features a whole section on content moderation, which discloses actions and metrics relating to violations of Ask.fm's Community Guidelines, including the prohibition of terrorist organisations and content. However, the diagrams depicting the metrics are broken, so the only way to determine whether there is any information on TVEC is unavailable. See https://about.ask.fm/legal/en/transparency.html
8. What information/fields of data are included in the TRs?	Ask.fm's transparency report includes: - The number of processed reports - The percentage of banned users, by category of policy violation (the diagram showing this is broken) - The top grounds for user ban, by country (the diagram showing this is broken) - The number of reports of violating texts - The number of reports for violating texts which were cleared (it is not explained what 'cleared' means) - The number of proactive detections by automated tools - The pre-moderation type of pattern results (the diagram showing this is broken)

	- The number of cleared media content (it is not explained what 'cleared' means) - The hash list hits during the reporting period (the diagram showing this is broken) With regard to TVEC in particular, Ask.fm informs that they reported 3 threats of extremism to law enforcement during 2020.
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 an annual basis.
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[177])

30. 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 of Terrorism on the platform, however, Vimeo prohibits: - any content that promotes or supports terror or hate groups. Art 5.2 (Content Terms) of its Terms of Service and Art 1.3 (Content Restrictions) of Vimeo's Acceptable Use Community Guidelines; - any depiction of unlawful acts or extreme violence and provision of instructions on how to assemble explosive/incendiary devices or homemade /improvised firearms. Art 5.2 (Content Terms) of Vimeo's Terms of Service and Art 1.3 (Content Restrictions) of Vimeo's Acceptable Use Community Guidelines; - the use of Vimeo by members of terror or hate group and gangs. Art 5.5 (Restricted Users) of Vimeo's Terms of Service and Art 2 (Restricted Users) of Vimeo's Acceptable Use Community Guidelines; - Hateful or discriminatory speech. Art 1.3 (Content Restrictions) of Vimeo's Acceptable Use Community Guidelines;

	- the purchase any (...) software services if you are (a) located in a country that is subject to a U.S. Government embargo or has been designated by the U.S. Government as a terrorist-supporting country; or (b) listed on any U.S. Government list of restricted parties. Art 5.5 (Restricted Users) of Vimeo's Terms of Service; Vimeo defines hate groups as organisations that *"aim to spread propaganda designed to radicalise and recruit people or aid and abet attacks*". Art 1.3 (Content Restrictions) of Vimeo's Acceptable Use Community Guidelines, and that have adopted, based upon its statements, leaders or activities, a hateful ideology. Art 2 (Restricted Users) of Vimeo's Acceptable Use Community Guidelines; and more specifically: Vimeo defines a hateful ideology as a set of beliefs that malign a group based upon personal characteristics. For U.S.-based groups, Vimeo considers the Southern Poverty Law Center's designations of hate groups to be conclusive. For non-U.S. groups, Vimeo may consider governmental or non-governmental designations. The absence of a group from any list of designated hate groups is not considered evidence that the group is not a hate group. Vimeo defines a terror group as a group that seeks to use criminal acts intended or calculated to provoke a state of terror in the general public to achieve political or ideological goals. Vimeo deems the U.S. Federal Bureau of Investigation's list of domestic terror groups and the U.S. Department of State's list of foreign terror groups as conclusive, but not exhaustive. A gang means any organization that uses fear, intimidation, or violence to conduct or further illegal activities or goals. Vimeo may consult relevant national and foreign law enforcement lists to determine whether an entity constitutes a gang. Also, content violates Vimeo's anti-hate and anti-discrimination section in Vimeo's Acceptable Use Community Guidelines in when it: ● Contains hateful or discriminatory speech;

	• Depicts (1) unlawful real-world acts of extreme violence, (2) vivid, realistic, or particularly graphic acts of violence and brutality, (...) and incite to violence • Violates any applicable law. all set out in Art 5.2 (Content Terms) of Vimeo'sTerms of Service; as well as, • Use or export any of Vimeo's services in violation of any U.S. export control laws; • Engage in any unlawful activity; all set out in Art 5.3 (Content Terms) of Vimeo's Terms of Service. and (1) is directed to a group based upon personal characteristics, such as race, color, national origina, ethnicity, religion, gender identity, and sexual orientation, disability and age; (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[1]). Content will generally be considered categorical hate speech if it: • Advocates for or celebrates violence against an individual or group based upon personal characteristics • Advocates or celebrates genocide • Calls for segregation or exclusion • Denies that certain historical events occurred (e.g., Holocaust denial)

	• Insults a minority group using a slur or "dog-whistle" code • Equates people to animals, filth, vermin, sexual predators, or criminals based upon personal characteristics • Spreads racial superiority theories or views • Spreads conspiracy theories about specific groups who share personal characteristics • Portrays a symbol of hate for no valid purpose all in Art 1.3 (Content Restrictions) of Vimeo's Acceptable Use Community Guidelines.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://vimeo.com/terms and https://vimeo.com/help/guidelines and also in the Legal FAQs: Guidelines Violations – Vimeo Help Center
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: (a) notifications of removals or (b) 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, Vimeo 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[1]). As a rule, Vimeo's 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.[2]).
4.1 Notifications of removals or other enforcement decisions	Enforcement decisions are notified via email except in cases of CSAM, extremist content, fraud or other illegal activities in which case Vimeo issues no notification.
4.2 Appeal processes against removals or other enforcement decisions	See Vimeo Law Enforcement Guidelines: Law Enforcement Guidelines on Vimeo and Art 5.3 (Appeals) of Vimeo's Acceptable Use Community Guidelines Account moderation decisions may be appealed (within 30 days of removal of content or 60 days after removal of the account), by completing a form. In the Form the user must (1) identify the content that was removed (and the URL if available); and (2) provide an explanation of why the user believes the decision is in error. Vimeo endeavours to respond within 30 days. If Vimeo finds good cause to reverse its initial decision, it will either restore the materials (if it still has them) or allow the user to resubmit them. Materials may not be re-uploaded pending an appeal. Vimeo reserves the right not to allow appeals in cases of extreme content, such as CSAM.
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. Moderators review these reports and take action accordingly. Art 5.1 (Approach to Moderation) of Vimeo's Acceptable Use Community Guidelines: "We endeavor to review specific content that is flagged by our users, third parties, and certain software-based systems. We do not endeavor to review every piece of content uploaded to our systems. In addition, when we do review content, it is usually for a particular reason, and so we do not endeavor to review it for all possible terms violations. Nor do we "pre-clear" any content before submission".

	Vimeo states that it may monitor users' accounts, content, and conduct, regardless of their privacy settings. Vimeo uses 'software-based systems' to flag violating content. Vimeo has signed an agreement with Active Fence to help identify TVEC content and have been working with them for years. Vimeo is not a GIFCT member (although it is currently applying) and does not participate in the 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. Vimeo is currently preparing one, however, and aims to publish it within the next year.
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?	According to Vimeo, it has happened on rare occasions.

31. 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 Rules have explicit prohibitions against: Threats of violence and incitement: Medium does not allow content or actions that threaten, encourage, or incite violence against anyone, directly or indirectly. Hateful content: Medium does not allow content that constitutes or promotes violence, harassment, or hatred against people based on characteristics like race, ethnicity,

	national origin, religion, caste, disability, disease, age, sexual orientation, gender, or gender identity. Medium does not allow posts or accounts that glorify, celebrate, downplay, or trivialise violence, suffering, abuse, or deaths of individuals or groups. This includes the use of scientific or pseudoscientific claims or misleading statistics to pathologise, dehumanise, or disempower others. Medium does not allow calls for intolerance, exclusion, or segregation based on protected characteristics, nor does it allow the glorification of groups which do any of the above. Lastly, Medium does not allow hateful text, images, symbols, or other content, including in usernames, profiles, or bios.
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.[178])

	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.[178])
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. Medium is not a member of the GIFCT, and does not participate in the 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.[178]).
7. Does the service issue transparency reports (TRs) on TVEC?	No. Medium issued a TR in 2015 (Medium, 2015[179]) 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.

32. 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, 2020[180]). 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, 2020[180]).
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.[181]) 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). LINE is not a member of the GIFCT, and does not participate in the 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. LINE has issued 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, 2020[180]). However, the last report (covering H1 2019) was issued in 2020, and contained no specific information on TVEC.
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 reported 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.

33. Picsart

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, Picsart's Community Guidelines contain the following explicit prohibitions: **Dangerous organisations and individuals**: Picsart cannot be used by groups or individuals that promote violence, hate, terrorism, crime, or other harmful behaviour. Picsart removes content and terminate accounts affiliated with gangs, terrorist organisations, cult communities, organised crime, and other violent or extremist groups. This includes content: ○ Depicting hand signals representing gang affiliation. ○ Depicting names, flags, slogans, monikers, logos, and other identifiers associated with such groups. ○ That glorifies or praises leaders or members associated with these groups. **Violence**: Picsart is not a place for graphic violence. Users may not post violent content involving humans or animals, including illustrated or computer-generated content, that is excessively bloody, vivid, gruesome, gory, or shocking. This includes content: ○ Depicting disturbing footage of war, car crashes and other accidents. ○ Depicting decapitations, suicide, terrorism, murder, or executions. ○ Depicting wounds where the injury is the central focus. ○ Depicting the torture, skinning, slaughter, mutilation, cruelty towards, or harm of animals or humans. ○ Depicting weapons with violent intent, including weapons positioned at another, weapons with blood or gore, or weapons widely associated with mass violence or dangerous events.

	○ Glorifying, commending, or idolising perpetrators of violence or violent events. **Hate**: Users may not post content that discriminates against, attacks, or promotes or incites hatred, harm, exploitation of, or bigotry or violence towards, individuals or groups based on following attributes: ○ Race. ○ Ethnicity. ○ Ancestry. ○ National origin or immigration status. ○ Religious affiliation. ○ Caste. ○ Gender. ○ Gender identity. ○ Sexual orientation. ○ Age. ○ Disability (physical or mental). ○ Disease. This includes content: ○ Depicting logos, symbols, flags, slurs, negative stereotypes, uniforms, salutes, gestures, caricatures, illustrations, or individuals related to hateful ideologies. ○ Condoning, idolising, or trivialising violent events that have occurred or may occur involving any of the attributes listed above. ○ Denying well-documented factual events have taken place or portraying such events as hoaxes or conspiracy theories. ○ Dehumanising or degrading individuals or groups based on the attributes listed above. ○ Justifying or promoting exclusions or segregation of individuals or groups based on the attributes listed above.

	○ Reinforcing harmful or negative stereotypes.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://picsart.com/terms-of-use and https://picsart.com/community-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?	Picsart broadly states that its Community Guidelines describe the type of behaviour and content that is prohibited on Picsart, and its team thoroughly investigates all reports of violations. Picsart removes content that violates its Community Guidelines and restricts or bans accounts with severe or repeated violations (Picsart, n.d.[182]). Determining whether there has been a violation of Picsart's Community Guidelines can be very nuanced, so Picsart reserves the right to make decisions it considers appropriate for the Picsart community.
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 Picsart. All reports are reviewed by Picsart's team of moderators (Picsart, 2015[183]). Picsart uses artificial intelligence in its content moderation efforts (Liao, 2019[184]) Picsart is not a member of the GIFCT, and does not participate in the GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Picsart removes content that violates its Community Guidelines and restricts or terminates accounts with severe or repeated violations. In certain circumstances, it may also report an account to the relevant authorities or 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?	Unknown.

34. 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 prohibit the use of Discord for the organisation, promotion or support of violent extremism. Discord considers violent extremism to be the support, encouragement, promotion, or organisation of violent acts or ideologies that advocate for the destruction of society, often by blaming certain individuals or groups and calling for violence against them (Discord, 2021[35]). Examples include racially motivated violent groups, religiously motivated groups dedicated to violence, and incel groups. Discord's Community Guidelines also 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.
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 that 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. Trust & Safety considers the severity of harm from the violative content, the potential for future harm on and off the platform, and whether an individual has grown and learned from their time away.
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 Discord's ToS and Guidelines. Discord has stated that, although it does not read users' private messages, it does investigate and take immediate appropriate action against any reported ToS violation by a server (something akin to a group or community under a common theme) or user (Liao, 2018[185]). 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 also relies on user moderators who are in charge of Discord's different communities. Discord recently created the Discord Moderator Academy (DMA), a comprehensive collection of resources intended to empower moderators to lead more effectively, manage teams, and learn more about the tools needed to help foster their communities. Those who pass the DMA Exam are eligible to apply to join Discord's moderator ecosystem (Discord, 2021[186]). Discord uses proactive tooling and resources to ensure that violent and hateful groups do not find a home on Discord. One recent step Discord has taken to improving safety on Discord was the July 2021 acquisition of an AI-based software company called Sentropy. The addition of this team is expected to allow Discord to expand its ability to detect and remove bad content, including hate, violence, and other forms of harm (Discord, 2021[187])

	With regard to violent extremism in particular, Discord states that its Trust & Safety team works to proactively find and remove servers and users engaging in high-harm activity like violent extremist organising. That team has developed frameworks based on academic research on violent extremist radicalisation and behaviour to better identify extremist users who try to use Discord to recruit or organise. In particular, Discord notes that violent extremism is nuanced and the ideologies and tactics behind them evolve fast. Thus, it does not try to apply its own labels or identify a certain 'type' of extremism. Instead, Discord evaluates user accounts, servers, and content that is flagged to them based on common characteristics and patterns of behaviour, such as: • Individual accounts, servers, or organised hate groups promote or embrace radical and dangerous ideas that are intended to cause or lead to real-world violence • These accounts, servers, or groups target other groups or individuals who they perceive as enemies of their community, usually based on a sensitive attribute. • They do not allow opinions or ideas opposing their ideologies to be expressed or accepted. • They express a desire to recruit others who are like them or believe in the same things to their communities and cause. Discord notes that the presence of one or two of these signals does not automatically mean that it would classify a server as 'violent extremist'. Whilst Discord might use these signs to determine a user or space's intent or purpose, Discord always wants to understand the context in which user content is posted before taking any action (Discord, 2021[35]). Discord is a member of the GIFCT, but does not participate in the 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. Discord’s last transparency reports (covering the periods July – December 2020 and January – June 2021) contain specific metrics on violent extremist content removal (Discord, 2021[188]) (Discord, 2021[187]).
8. What information/fields of data are included in the TRs?	Discord’s last transparency report discloses: - the overall number of reports received, as well as the number and percentage that fell within each prohibited category (one of which is extremist or violent content) - the number and percentage of the reports on which Discord took action, and the report action rate, by prohibited category - The total number of account deletions (excluding spam), by prohibited categories - The total number of server deletions, by prohibited categories - The number of accounts warned, accounts deleted after warning, servers warned, and servers deleted after warning, by prohibited categories - The number of server deletions proactively deleted and reactively deleted, by prohibited categories - The percentage of accounts reinstated on appeal, by prohibited categories. Discord also informs that in the weeks before the storming of the US Capitol, its team of counter-extremism experts began monitoring the situation, and proactively removed a number of servers involved in discussing and organising the event in December 2020. The team removed 27 servers and 857 accounts the day of 6 January 2021 (Discord, 2021[187]).
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 semi-annual basis.
11. Has this service been used to post TVEC?	Yes. See Section 8 above.

35. Twitch

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, Twitch's Community Guide provide that 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. Users may not display or link terrorist or extremist propaganda, including graphic pictures or footage of terrorist or extremist violence, even for the purposes of denouncing such content. Moreover, 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, encourages, or facilitates violence, among other things, based on race, ethnicity, national origin, religion, sex, gender, gender identity, sexual orientation, age, disability, medical condition, physical characteristics, or veteran status. Twitch indicates that the following is considered hateful conduct: • Promoting, glorifying, threatening, or advocating violence, physical harm, or death against individual(s) or groups on the basis of a protected characteristic, including age. • Using hateful slurs, either untargeted or directed towards another individual.

	• Posting, uploading, or otherwise sharing hateful images or symbols, including symbols of established hate groups and Nazi-related imagery. • Speech, imagery, or emote combinations that dehumanise or perpetuate negative stereotypes and/or memes. • Content that expresses inferiority based on a protected characteristic, including, but not limited to, statements related to physical, mental, and moral deficiencies. • Calls for subjugation, segregation or exclusion, including political, economic, and social exclusion/segregation, based on a protected characteristic, including age. • Content that encourages or supports the political or economic dominance of any race, ethnicity, or religious group, including support for white supremacist/nationalist ideologies. • Expressions of contempt, hatred, or disgust based on a protected characteristic. • Mocking the event/victims or denying the occurrence of well-documented hate crimes, or denying the existence of documented acts of mass murder/genocide against a protected group. • Content that makes unfounded claims assigning blame to a protected group, or that otherwise intends to incite fear about a protected group as it relates to health and safety. • Encouraging the use of or generally endorsing sexual orientation conversion therapy. • Membership, support, or promotion of a hate group, including sharing hate group propaganda materials. • Creating accounts dedicated to hate, such as through abusive usernames.
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/article/about-account-suspensions-dmca-suspensions-and-chat-bans?language=en_US

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?	Twitch takes enforcement action against accounts that violate its ToS and/or Community Guidelines. Twitch considers several factors when reviewing reports of violations, including the intent and context, the 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 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 with no opportunity to appeal. Twitch also notes that in exceptional circumstances, I may pre-emptively suspend accounts when it believes an individual's use of Twitch poses a high likelihood of inciting violence. In weighing the risk of harm, Twitch considers an individual's influence, the level of recklessness in their past behaviours (regardless of whether any past behaviour occurred on Twitch), whether or not there continues to be a risk of harm, and the scale of ongoing threats. Lastly, Twitch notes that it enforces against severe offenses committed by members of the Twitch community that occur outside its services, such as hate group membership, terrorist recruitment, sexual assault, and child grooming. Twitch investigates reports that include verifiable evidence

	of these behaviours and, if it is able to confirm, issue enforcements against the relevant users.
4.1 Notifications of removals or other enforcement decisions	There are warnings, depending on the nature of the violation.
4.2 Appeal processes against removals or other enforcement decisions	In cases not resulting in immediate suspension, 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 explains a layered approach to safety on its platform. The foundation is its Community Guidelines, which are clear enough to set the boundaries as to what is and is not allowed on Twitch. Then there is service level safety. This is composed of three parts: machine detection, user reporting and review and enforcement. Machine Detection: Over the last two years, Twitch has implemented 'machine detection' technologies that scan content on the service to remove harmful or inappropriate content, or flag it for review by human specialists. Examples of this are nudity, sexual content, gore and extreme violence. Twitch is predominantly a live-streaming service, and most of the content that is streamed is not recorded or uploaded. Because content is viewed as it is created, live-streaming provides a particularly challenging environment for machine detection to keep up. Nevertheless, Twitch has found ways to use machine detection to bolster proactive moderation on Twitch, and it will continue to invest in these technologies to improve them. User Reporting: Community reports are a crucial part of maintaining the safety and trust of Twitch's community and upholding its Community Guidelines. User reporting is particularly effective on Twitch because the vast majority of the content on Twitch - video and chat - is public. Twitch encourages creators, moderators, and viewers to report content that violates its Community Guidelines so Twitch can take appropriate service-wide action. User reports are sent to Twitch's team of content moderation professionals to review. Review and Enforcement: There is a group of highly trained and experienced professionals who review user reports,

	and content that is flagged by Twitch's machine detection tools. These content moderation professionals work across multiple locations, and support over 20 languages, in order to provide 24/7/365 capacity to review reports as they come in across the globe. Reports are prioritised so that the most harmful behaviour can be dealt with most quickly. Review time for any given report is dependent on a number of factors including the severity of the report, the availability of evidence to support the report, and the current volume of the report queue. Twitch also employs a team of experienced investigators to delve into the most egregious reports, and works with law enforcement as necessary. Then, there is channel-level safety, in charge of the channel creator. Twitch enable creators to set their own standards of acceptable and unacceptable community behaviour, with Twitch's Community Guidelines providing a baseline standard that all communities are required to uphold. To foster a culture of accountability, creators can leverage other members of their community and create a team of moderators, who assist the creator by moderating chat in the creator's channel (moderators can be easily identified in chat by the green sword icon that appears next to their username). Many Twitch creators ask trusted members of their communities to help moderate chat in the creator's channel. These channel moderators ("mods") and moderation tools are the foundation of chat moderation in every creator's Twitch channel. Mods play many roles, from welcoming new viewers to the channel, to answering questions, to modelling and enforcing community standards. Twitch provides both creators and their moderators with a powerful suite of tools such as AutoMod, Chat Modes, and Mod View to make their roles as easy and intuitive as possible. These tools provide the ability to automatically filter chat, allow creators and moderators to see (and delete) questionable chat messages before they are displayed on the channel, give users "time outs" (lock them out of chat for a period of time) or permanently block them from the channel. Twitch's suite of moderation tools supports two objectives: identifying potentially harmful content for moderator review, and scaling moderator controls to support fast-moving Twitch chat messages. (Twitch, 2021[49]). 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	Depending on the nature of the violation, Twitch takes a range of actions including issuing a warning, a temporary suspension (1-30 days), and for the most serious offenses, an indefinite suspension from Twitch. If any content that contains the violation has been recorded on the service, Twitch will remove it.
7. Does the service issue transparency reports (TRs) on TVEC?	Yes (Twitch, 2021[49]).
8. What information/fields of data are included in the TRs?	Twitch explains that it is a live-streaming service, and the vast majority of the content on Twitch is ephemeral. For this reason, it 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 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's transparency report focuses on enforcements mostly. Twitch's last transparency report includes the following metrics: - Number of user reports for all types of violations during the reporting period (H1 and H2 2020); - Total number of enforcement actions; - Number of enforcement actions for reports of hateful conduct, sexual harassment and harassment, and number of enforcement actions for reports of these violations per thousand hours watched; - Number of enforcement actions for reports of violence, gore, threats and other shocking content, and number of enforcement actions for reports of these violations per thousand hours watched; - Number of enforcement actions for reports of terrorism, terrorist propaganda and recruitment, and number of enforcement actions for reports of these violations per thousand hours watched. Twitch notes that in 2020 they did not have any instances of live-streamed terrorist activity. The

	enforcements issued in this category were for showing terrorist propaganda (77 enforcements in 2020), and for glorifying or advocating acts of terrorism, extreme violence or large-scale property destruction (10 enforcements in 2020). - Number of enforcement actions for reports of adult nudity, pornography and sexual conduct, and number of enforcement actions for reports of these violations per thousand hours watched; - Number of enforcement actions for reports of spam and other community guidelines violations, and number of enforcement actions for reports of these violations per thousand hours watched - The percentage of hours of live content watched in channels that had (i) automod enabled, (ii) at least one active moderator, and (iii) both; - The number of proactive (Blocked Terms and Automod) and manual (Mods) removals of chat messages - The number of channel enforcement actions, broken down by Timeouts and Channel bans.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	No specific information provided.
10. Frequency/timing with which TRs are issued	On an annual basis.
11. Has this service been used to post TVEC?	Yes. See section 8 above.

36. Likee

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, Likee's Community Guidelines contain the following prohibitions: **Dangerous individuals and organisations**: Likee prohibits individuals and organisations from using Likee to promote terrorism, crime, and other behaviours that pose a serious danger to society. Likee will ban accounts that threaten or endanger users or public safety. Likee also prohibits all content involving terrorist behaviour. Specifically, any content involving the following is prohibited: hate groups, violent

	extremist organisations, homicides, human trafficking, organ trafficking, arms trafficking, drug trafficking, kidnapping, extortion, blackmail, money laundering, fraud, and cybercrime. Likee will ban the accounts of terrorists, terrorist organisations, and criminals. The following content is prohibited: • Content that includes the names, symbols, signs, flags, slogans, uniforms, gestures, portraits, or other items representing dangerous individuals and/or organisations • Content that praises, glorifies, or supports dangerous individuals and/or organisations • Content involving violent harm to personal safety, such as assaults or kidnapping • Content that may endanger the personal safety of others, such as sneak attacks • Content involving the purchase, sale, or exchange of illegally obtained goods • Content that provides instructions for criminal activities • Other crime-related content **Hate Speech** Likee does not allow hate speech to be posted or disseminated on its platform. The following content is prohibited: • Content involving racial discrimination • Content that incites religious hatred • Content that promotes fascism • Any language or action that promotes or gives evidence to the rejection, isolation, or discrimination against an individual **Violence and Violent Images** Content involving behaviour that can lead to the death of a victim or threats of violence in any form is prohibited. Likee also prohibits frightening content, especially content that promotes or glorifies violence and violent images. Content that may lead to violent acts or content that involves violence of any other form is prohibited, including but not limited to:

	• Content involving the intention to commit highly violent acts • Content containing incitement to violence • Content describing the violent or accidental death of a real person • Content describing physical violence
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://mobile.like.video/live/about/userAgreement , https://likee.video/live/page-about/user-agreement.html and https://mobile.likee.video/live/page-about/community.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?	Likee broadly states that it may, but will not have any obligation to, review, monitor, display, reject, refuse to post, store, maintain, accept or remove any content posted by the user, and it may, in it sole discretion, delete, move, re-format, remove or refuse to post or otherwise make use of the content without notice or any liability to the user or any third party in connection with its operation of Likee in an appropriate manner. In particular, Likee may do so to address content that comes to its attention that it believes is offensive, obscene, violent, harassing, threatening, abusive, illegal or otherwise objectionable or inappropriate.
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)	Likee has a team of moderators proactively detecting violating content and reviewing user reports. Likee is not a member of the GIFCT, and does not participate in the 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, Likee may remove the violating content, terminate or suspend the infringer's access to Likee, and refer matters 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?	Unknown.

37. 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[172]) that, for the purposes of its services, terrorist content is material posted by or in support of organisations included on the Consolidated United Nations Security Council Sanctions List (United Nations Security Council, n.d.[173]) that depicts graphic violence, encourages violent action, endorses a terrorist organisation 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 organisations. 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.
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 deploys a variety of scanning technology, artificial intelligence, external partnerships, and human moderation operations solutions to detect and investigate TVEC. Furthermore, users are able to report content that violates Skype's ToS or is otherwise unlawful or objectionable. Moderators review the reports 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[172]).

	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, 2020-2021[175]). 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 TVEC actioned reported by users or third parties • % 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	On a semi-annual basis.
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, n.d.[189]).

38. 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. Also, making threats of violence or spreading hate speech as well as victimising or belittling an individual or group of people based on religion, culture, race, ethnicity, nationality, sexual or gender identity, developmental differences, illness, etc., is forbidden. 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 content provided for in Russian law.
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?	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?	VK immediately deletes content that ill-intentioned users distribute in order to spread ideologies of hate or hostility. If a user seriously or repeatedly violates the platform's standards, their account or community will be permanently blocked. VK notes that hate speech can vary in how it looks, and it always takes context into account. VK's moderators look for signs in the content that confirm that the user posting it was doing so maliciously, such as:

	• animosity based on certain characteristics or differences. • offensive behaviour, contempt toward other people's values or views. • expression of personal superiority, accompanied by a baseless and unfair attitude toward a specific individual or group of people. BK blocks accounts and communities that spread content containing: • direct threats to life or well-being (for example, language threatening someone's "destruction"). • calls to suicide in any form (suggestions or insistence to do something with oneself as well as descriptions of suicide methods). • hostile, threatening or violence-inducing language, attacks on a person or group of people with the goal of degrading human dignity or claiming their inferiority (such content may contain a link to a profile, person's photo, their home address or phone number, along with promises to "find" them, insults or phrases such as "you know what to do"). • calls for isolation or segregation (for example, content suggesting that certain people need to live somewhere else: "take them away", "let them live among themselves in… and not stand out", and so on), wishes of serious harm or calls to inflict it, encouragement of victimisation or offensive behaviour, masked calls or incitement to violence (often followed by calls to drive certain people out and so on). • verbal assertion of superiority of some groups over others to rationalise violence, discrimination, segregation, or isolation on the basis of religion, ethnicity, nationality, sexual or gender identity, developmental differences or illness (for example, this may be done by comparing a specific group of people to insects, filth, sub-humans, inferior types, and other such language).
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 content being deleted or blocked, 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[190]). 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. VK is not a member of the GIFCT, and does not participate in the 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, 2021[42]) and (VK, 2021[191]).
8. What information/fields of data are included in the TRs?	Number of pieces of content, profiles and communities blocked due to promotion of hatred or hostility and drug advocacy or distribution (Q1 and Q2 2020 statistics).
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[192]).

39. Xigua Video

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 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)	Users can report any type of unlawful activity or content on Xigua. Whilst Xigua's content moderation practices are kept in secret, former ByteDance employees have disclosed widespread use of moderators and automated tools to filter content and detect 'problematic' speech (Lu, 2021[59]). Xigua is not a member of the GIFCT, and does not participate in the 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.

40. 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.[193]). Users can report content that violates Odnoklassniki's ToS. Odnoklassniki's team reviews such reports and decides what actions to take. 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. On 6 July 2020 Odnoklassniki presented Robbie, an automated platform for the analysis of content based on neuronets and big data technology. According to Odnoklassniki, the platform developed by a command of social network will help to build processes on processing of content without involvement of additional personnel (Dusaleev, 2020[194]). Odnoklassniki is not a member of the GIFCT, and does not participate in the 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 (Powell, 2019[195]).

41. 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 Community Guidelines prohibit the posting of content that is illegal or prohibited, such as 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	Photo moderation decisions are notified (see section 5 below).
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. Flickr users are under the obligation to moderate their own content according to Flickr's safety levels. In this sense, Flickr relies on user moderators to moderate content. Flickr notes that the use of auto-moderation technology helps in the efforts to ensure all content is moderated properly. The Moderation Bot detects content from uploads and

	automatically updates mis-moderated content to the correct moderation level according to Flickr's established safety levels. When this occurs, the user will receive a private notification under the bell icon that lets the user know about the mismatch and directs them to the photo in question. (Flickr, 2021[196]). Flickr is not a member of the GIFCT, and does not participate in the 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[197]).

42. 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. Whilst Huoshan's content moderation practices are kept in secret, former ByteDance employees have disclosed widespread use of moderators and automated tools to filter content and detect 'problematic' speech (Lu, 2021[59]). Huoshan is not a member of the GIFCT, and does not participate in the 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.

43. Kakao

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, Kakao's ToS prohibit violent content and behaviour that enables or motivates illegal activities. Also, Kakao prohibits all forms of discrimination which promotes stereotypes based on region, disability, race, ethnicity, gender, age, job and religion.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.kakao.com/en/terms and https://www.kakao.com/policy/oppolicy?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?	Kakao 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 (Kakao, n.d.[198]).
4.2 Appeal processes against removals or other enforcement decisions	Users can appeal the actions taken, and Kakao 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 Kakao's policies. In addition, users can report any content that violates Kakao's policies. Kakao'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, Kakao 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. Kakao is not a member of the GIFCT, and does not participate in the GIFCT's Hash Sharing Consortium.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	In case of violations of Kakao's policies, Kakao 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. Kakao, however, does issue transparency reports (Daum Kakao, 2020[199]) 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	Not applicable.
11. Has this service been used to post TVEC?	Unknown.

44. 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. Smule is not a member of the GIFCT, and does not participate in the 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.

45. 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 Etiquette Policy provide that commentaries that are overly aggressive, personally insulting or needlessly abusive are prohibited ('Prohibited Commentaries'). Users must refrain from comments which are racist, bigoted, or which otherwise offensively target a philosophy or religion. In addition, users must avoid making offensive remarks based on gender or sexual preference, as well as comments or critique which is intended to be a direct insult to an individual, group, or genre of artwork. Hate propaganda is met with zero tolerance. 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	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

there notifications of removals or other enforcement decisions and appeal processes against them?	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[200]). 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.[201]). 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.[202]). 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.[203]).
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 after having a reactive moderation system for decades, they are now implementing a new moderation system that is a proactive one. DeviantArt's team is aiming to remove deviations that violate site policies before they are reported, while still allowing deviants to report deviations. This approach is enabled by a new technology that helps DeviantArt to better identify deviations in violation of its policies. This technology is intended to aid the moderation team work through deviations more rapidly and efficiently — the moderation team itself is not being replaced, nor will deviations have actions automatically taken by Artificial Intelligence. The new technology serves to help the moderation team take a more proactive approach to reviewing deviations, and they still handle all moderation decisions (DeviantArt, 2020[204]). DeviantArt is not a member of the GIFCT, and does not participate in the 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, 2019[205]).

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.[206]), 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 marginalisation. *Terrorist activities*: Google does not permit terrorist organisations 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 organisations' 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.[207]).
4.1 Notifications of removals or other enforcement decisions	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. 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.
4.2 Appeal processes against removals or other enforcement decisions	If a file has a violation notice, the owner can request a review of the violation.
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[208]). 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[209]). GoogleDrive is not a member of the GIFCT, and does not participate in the 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.[210]) - Report illegal materials to appropriate law enforcement authorities
7. Does the service issue transparency reports (TRs) on TVEC?	No. Google issues TRs (Google, n.d.[211]) 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, 2018[212]).

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 will suspend or terminate the user's access.
4.1 Notifications of removals or other enforcement decisions	In cases of breaches of its ToW, Dropbox provides reasonable advance notice via the email address associated with the user's account and gives the user 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.
4.2 Appeal processes against removals or other enforcement decisions	Appeals against content takedowns, including TVEC, are allowed (Volkmer, 2019[213]).
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 Dropbox's ToS and policies. Dropbox's team reviews these reports, investigates the alleged violation, and takes appropriate action. 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.[214]). In 2018 Dropbox tested and then implemented a trusted flagger program. This enables Dropbox to prioritise content removal referrals from organisations, such as countries' internet referral

	units, once a high degree of accuracy that the content they refer is harmful is established. Dropbox has also entered into URL sharing agreements with social media companies, including Twitter, in order to prioritise removal of material hosted on Dropbox that has been widely shared on their platforms. In addition, Dropbox participates in the EU Internet Forum to discuss ways to reduce the spread of terrorist content with European policymakers, along with other public and private sector organisations addressing this challenge (Volkmer, 2019[213]). 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 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.[215]) 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[151]).

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 organisations included on the Consolidated United Nations Security Council Sanctions List (United Nations Security Council, n.d.[173]) 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[172]).
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 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?	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[172]).
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 deploys a variety of scanning technology, artificial intelligence, external partnerships, and human moderation operations solutions to detect and investigate TVEC. Furthermore, users are able to report content that violates Microsoft's policies. Moderators review the reports and decide on the best action to implement. 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: With regard to TVEC in particular, Microsoft has informed the following: "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." (Microsoft, 2016[172])
7. Does the service issue transparency reports (TRs) on TVEC?	Yes. TVEC numbers for OneDrive are included in Microsoft's Digital Safety Content Report (Microsoft, 2020-2021[175]). 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 TVEC actioned reported by users or third parties

	• % 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	On a semi-annual basis.
11. Has this service been used to post TVEC?	Yes. ISIS videos have been hosted on OneDrive (Counter Extremism Project, 2018[216]).

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. However. 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[217]), 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.[218]). Genuine calls to violence are also prohibited. This includes 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. Content that glorifies acts of violence or its perpetrators is removed.
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	WordPress.com has worked in conjunction with experts on online extremism, as well as law enforcement, to develop policies to

Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	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 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 Trust & 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[219]).
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.[220]).

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 moderator 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. WordPress.com is a member of the GIFCT.
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.[221]). The last transparency report included data from 1 July to 31 December 2021.
8. What information/fields of data are included in the TRs?	- Number of IRU extremist (not specifically violent extremist) content notices - Number of notices where sites/content were removed as a result - Percentage of notices where 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 (e.g. Europol) or country. Also, in the Summary section of its transparency reports, Automattic reports the number of sites/content specified in the IRU notices for the period between 1 January 2018 – 31 December 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 - 2020: 1 Jul – 31 Dec - 2021: 1 Jan – 30 Jun - 2021: 1 Jul – 31 Dec
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[222]) and oversight (Wikipedia, 2020[223]) 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[224]). 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[225]). 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[226]). The deletion process encompasses the processes involved in implementing and recording the community's decisions to delete pages and media (Wikipedia, 2020[227]). 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[228]). 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[229]). Wikipedia is not a member of the GIFCT, and does not participate in the 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.[230]) 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 - The Global Top-50 TVEC-intensive Services

Mainstream TVEC-intensive Services						
Rank	Name of service (parent company)	Type of service	No. of URLs linking to TVEC	TVEC categories included in links (Jihadi, far-right)	Issues TVEC transparency reports	Provided feedback / comments on its profile
1	Telegram (Telegram Messenger LLP)	Messaging app	2,361,948	Both	N	N
2	YouTube (Alphabet, Inc.)	Video streaming platform	1,583,412	Both	Y	N
3	Twitter (Twitter, Inc.)	Short messages-focused social networking platform	850,888	Both	Y	N
4	Facebook (Meta, Inc.)	Social networking platform	152,435	Both	Y	N
5	Instagram (Meta, Inc.)	Social networking platform	73,905	Both	Y	N
6	TikTok (ByteDance Technology Co.)	Short video app	36,337	Both	Y	Y
7	VK (Mail.Ru Group)	Social networking platform	36,334	Far-right	N	N
8	WhatsApp (Meta, Inc.)	Messaging app	17,895	Both	N	N
9	Element (New Vector Ltd)	Messaging app	13,766	Both	N	Y
10	Discord (Discord, Inc.)	Chat platform	4,760	Far-right	Y	N

File-sharing TVEC-intensive Services						
Rank	Name of service (parent company)	Type of service	No. of URLs linking to TVEC	TVEC categories included in links	Issues TVEC transparency reports	Provided feedback / comments on its profile
11	Telegraph (Telegram Messenger LLP)	Anonymous blogging platform	25,427	Both	N	N
12	Archive.org (The Internet Archive, a 501(c)(3) non-profit internet library)	Internet library	19,065	Both	N	N
13	Justpaste.it (Wise Web Mariusz Żurawek)	Anonymous pastebin site	9,669	Jihadi	Y	Y
14	Files.fm (Files.fm, SIA)	Cloud-based file sharing	9,146	Both	N	Y
15	Google Drive (Alphabet, Inc.)	Cloud-based file sharing	8,896	Both	N	N
16	Tlgur (Telegram Messenger LLP)	File sharing	8,231	Both	N	N
17	Dropbox (Dropbox, Inc.)	Cloud-based file sharing	7,663	Both	N	N
18	MediaFire (MediaFire, LLC)	Cloud-based file sharing	7,111	Both	N	N
19	Google Docs (Alphabet, Inc.)	Online word processor	6,362	Both	N	N
20	Mega.nz (Mega Ltd)	Cloud-based file sharing	6,032	Both	Y	N
21	Pixeldrain (Fornaxian Technologies)	File sharing	5,982	Both	N	N
22	Uploadgram (Telegram Messenger LLP)	Anonymous file sharing	5,731	Both	N	N
23	File.io (Mr Cowboy LLC)	Anonymous file sharing	5,444	Both	N	N

24	Gofile.io (Wojtek SAS)	File sharing	5,053	Both	N	N
25	Anonfiles (unknown)	Anonymous file sharing	4,734	Both	N	Y

Far-right-focused TVEC-intensive Services						
Rank	**Name of service (parent company)**	**Type of service**	**Visits**	**Unique visitors**	**Issues TVEC transparency reports**	**Provided feedback / comments on its profile**
26	Bitchute.com (Bit Chute Ltd)	Video streaming platform	315,120,431	77,161,929	N	N
27	Rumble.com (Rumble, Inc.)	Video streaming platform	281,700,069	118,231,593	N	Y
28	Gab.com (Gab AI, Inc.)	Social networking platform	185,829,025	31,379,454	N	N
29	Patriots.win (Patriots, LLC)	Social news aggregation, web content ranking and discussion website	139,453,406	5,286,808	N	N
30	Parler.com (Parler, Inc.)	Social networking platform	114,949,343	43,760,475	N	N
31	Odysee (Odysee, Inc.)	Video streaming platform	95,955,315	26,972,611	N	N
32	Brandnewtube (My Media World Ltd)	Video streaming platform	38,072,805	6,478,066	N	N
33	Gettr (Gettr USA, Inc.)	Social networking platform	32,481,188	5,973,673	N	N
34	8kun.top (N.T. Technology, Inc.)	Content-sharing and discussion website	31,595,372	8,166,601	N	N
35	Red Voice Media (Unknown)	News aggregation and discussion website	16,996,904	6,094,840	N	N

36	Thedonald.win (Jody Williams)	Social news aggregation and discussion website	13,469,271	1,834,145	N	N
37	WeGo Social (AnonUp LLC)	Social networking platform	13,255,017	330,122	N	N
38	SafeChat (SafeChat, Inc.)	Messaging app	11,181,357	1,543,158	N	N
39	88msn.com (Unknown)	Content-sharing and discussion platform	2,884,835	350,504	N	N
40	Doxbin.org (~kt & Brenton)	File sharing and publishing website	2,837,341	1,241,029	N	N
41	Wimkin.com (DreamTeam Development, LLC)	Social networking platform	2,298,980	630,793	N	N
42	Mzwnews.com (John De Nugent)	News blogging platform	1,706,533	217,595	N	N
43	Worldtruthvideos. website (Unknown)	Video streaming platform	699,672	100,745	N	N
44	Xephula (Unknown)	Social networking platform	667,431	147,950	N	N
45	Thegreaterrreset.org (Unknown)	Interest-based social networking platform	546,461	216,051	N	N

46	Nordfront.dk (Unknown)	Propaganda dissemination and recruitment website	532,376	138,254	N	N
47	Lookaheadamerica.org (Unknown)	Propaganda dissemination and recruitment website	256,490	181,464	N	N
48	Patriotfront.us (Unknown)	Propaganda dissemination and recruitment website	174,910	141,325	N	N
49	Vastarinta.com (Unknown)	Propaganda dissemination and recruitment website	141,692	82,983	N	N
50	Nordicresistance movement.org (Unknown)	Propaganda dissemination and recruitment website	80,221	49,441	N	N

Annex D - Profiles of the Top-50 TVEC-intensive Services

Mainstream TVEC-intensive Services

1. Telegram

See profile 13 in Annex B.

2. YouTube

See profile 2 in Annex B.

3. Twitter

See profile 21 in Annex B.

4. Facebook

See profile 1 in Annex B.

5. Instagram

See profile 6 in Annex B.

6. TikTok

See profile 10 in Annex B.

7. VK

See profile 38 in Annex B.

8. WhatsApp

See profile 4 in Annex B.

9. Element

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No specific definition is provided. Element's ToS prohibit use of the Element app and services in violation of any applicable laws or regulations related to the access to or use of Element; for any unlawful purposes or in support of illegal activities under UK/EU law. Harassment is defined in the matrix.org community rooms' Code of Conduct (Matrix is Element's parent company). This term includes: • Offensive comments related to gender, gender identity and expression, sexual orientation, disability, mental illness, neuro(a)typicality, physical appearance, body size, race, age, regional discrimination, political or religious affiliation • Threats of violence, both physical and psychological • Incitement of violence towards any individual, including encouraging a person to commit suicide or to engage in self-harm • Deliberate intimidation
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Element's ToS are available at https://element.io/terms-of-service and the Matrix Foundation's homeserver terms and conditions are available at https://matrix.org/legal/terms-and-conditions The Matrix community rooms' Code of Conduct is available at https://matrix.org/legal/code-of-conduct
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 Element, if someone has been harmed or offended, it is Element's responsibility to listen carefully and respectfully, and do its best to right the wrong. A formal appeals process for the matrix.org homeserver is currently under development and will be made public soon.
4.1 Notifications of removals or other enforcement decisions	No notifications are specified. However, if a user thinks Element removed their access by mistake, users can send

	an email to support@matrix.org, and Element will provide an explanation for the decision.
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 violating content via abuse@matrix.org. After filing a report, a representative will contact the complainant personally, review the incident, follow up with any additional questions, and make a decision as to how to respond. If the person who is harassing the complainant is part of the response team, they will recuse themselves from handling the incident. If the complaint originates from a member of the response team, it will be handled by a different member of the response team.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Breaches of Element's ToS enable Element to block, restrict, disable, suspend or terminate the infringer's access to all or part of Element.
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.

10. Discord

See profile 34 in Annex B.

File-sharing TVEC-intensive Services

11. Telegraph

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There are neither ToS nor Community Guidelines/Standards. Thus, there is neither a specific definition of nor a general prohibition on TVEC.

2. Manner in which the ToS or Community Guidelines/Standards are communicated	There are neither ToS nor Community Guidelines/Standards.
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?	Not applicable.
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)	No information available.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Not 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.

12. Archive.org

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No specific definition is provided. Archive.org's ToS provide that users may not to act in any way that might give rise to civil or criminal liability; not harass, threaten, or otherwise annoy anyone; and not act in any way that might be harmful to minors, including, without limitation, transmitting or facilitating the transmission of child pornography.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Archive.org's ToS are available at https://archive.org/about/terms.php
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?	No policies or 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 content that violates Archive,org's ToS via email with the URL (web address) of the item to info@archive.org.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Violation of Archive.org's ToS entitles Archive.org to immediately deactivate any password it has issued to the infringer and bar the infringer from accessing Archive.org's collection of materials.
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.

13. Justpaste.it

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No specific definition is provided. However, Justpaste's ToS provide that users may not post or upload any content which contains material which is unlawful for users to possess in the country in which they are resident, or which would be unlawful for JustPaste.it to use or possess in connection with the provision of its services (including Terrorism content). "Terrorist content" means any information the dissemination of which amounts to offences specified in Directive (EU) 2017/541 or terrorist offences specified in the law of a Member State concerned, including the dissemination of relevant information produced by or attributable to terrorist groups or entities included in the relevant lists established by the European Union or by the United Nations.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Justpaste.it's ToS are available at https://justpaste.it/terms
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?	If a user finds their Content or account blocked, but they are convinced that they have not broken any part of Justpaste.it's ToS, they can appeal to JustPaste.it via email at: support@justpaste.it. Only content created by a registered account can be appealed.
4.1 Notifications of removals or other enforcement decisions	Registered accounts get an email when the account is banned. There is no possibility to notify users about removal of content created as anonymous (without registered account).

4.2 Appeal processes against removals or other enforcement decisions	Appeal procedure is described in https://justpaste.it/terms/appeal (Terms 7.9)
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Abusive content that is against Terms of Service can be reported directly to JustPaste.it by email at: support@justpaste.it. To speedup process of content moderation the word "Abuse" in the message title must be included. In message body the full URL of each reported material in separate line must be provided. The justification for report can be included at the end of the message. Reports without any explanation of violation or not containing links to reported content may not be processed. Justpaste.it observes that it receives reports from governments and law enforcement agencies regarding content published on JustPaste.it that violate its ToS. Reported content is reviewed by Justpaste.it' staff against its ToS and Polish law before taking action (Justpaste.it, 2021[231]). As a member of Hash Sharing Consortium, Justpaste.it uses anonymous identifiers of content that were shared by other companies to detect abusive materials on its platform (Justpaste.it, 2020[232]).
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	On becoming aware of any potential violation of its ToS, JustPaste.it may remove violating content and/or terminate the accounts of the user found to be in violation of said terms.
7. Does the service issue transparency reports (TRs) on TVEC?	Yes. Justpaste.it published its first TR in 2020 (covering 2019). Its last report covers the year 2021 (available at https://justpaste.it/transparency_report_2021)
8. What information/fields of data are included in the TRs?	In its last TR (covering 2021), Justpaste.it reported the number of requests received from governments and law enforcement agencies regarding illegal content (broken down by EU requests, UK requests, Russian requests and Turkish requests), the percentage of such requests which related to terrorist content, and the percentage of terrorist content reports in which Justpaste.it took action and blocked the terrorist content. Justpaste.it observes that the content of its transparency report is based on Tech Against Terrorism's recommendation for Transparency Reporting on small platforms.

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 yearly basis.

14. Files.fm

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No specific definition is provided. However, Files.fm's ToS provide that users may not carry out and submit, send or share information (including photos, video and audio materials) that: • infringes personal dignity and respect; • incite violence, racial hatred, or other illegal activities; • is vulgar, libellous, or otherwise offensive; and • violates laws and regulations.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Files.fm's ToS are available at https://files.fm/terms
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?	Files.fm broadly states that it may review users' conduct and content to determine compliance with its ToS, although it has no obligation to do so.
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 may report violating content by submitting a content removal application ("Application"). The Application must contain the following information:

	• Full name of applicant, personal identification number, position, organization. • Contact information: Phone, E-mail and Address. • Links to the problematic content and at least one reference url, where the content is linked - accessible to the general public. • Description of the problem - what exactly it violates. Applications are processed as quickly as possible - a few days, but processing times may be longer. After moderators review the Application, they take any actions deemed appropriate (e.g. deleting or disabling content, and termination of account of repeat infringers).
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Files.fm may suspend or terminate a user's use of its services in cases of violation of its ToS, or where use of Files.fm is in a manner that could cause Files.fm legal liability, disrupt its services or disrupt others' use 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.

15. Google Drive

See profile 46 in Annex B.

16. Tlgur

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There are neither ToS nor Community Guidelines/Standards. Thus, there is neither a specific definition of nor a general prohibition on TVEC.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	There are neither ToS nor Community Guidelines/Standards.

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?	Not applicable.
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)	No information available.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Not 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.

17. Dropbox

See profile 47 in Annex B.

18. MediaFire

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No specific definition is provided. However, MediaFire's ToS prohibit the distribution of content that is libelous, defamatory, obscene, pornographic, abusive, harassing, threatening, unlawful or promotes or encourages illegal activity; as well as use of MediaFire's services for any illegal purpose.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	MediaFire's ToS are available at https://www.mediafire.com/policies/terms_of_service.php
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?	MediaFire broadly states that it reserves the right to determine what is harmful to its users, operations, or reputation including any activities that restrict or inhibit any other user from using and enjoying its services.
4.1 Notifications of removals or other enforcement decisions	When MediaFire removes or disables Content for policy violations, the user who posted the Content may receive a strike. The user is notified of the violation.
4.2 Appeal processes against removals or other enforcement decisions	If a user feels their account was suspended in error, they can contact MediaFire's support department with detailed information for further evaluation.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Complaints about violators of MediaFire's policies can be directed at its abuse department. Complaints are processed by MediaFire's Customer Support team. MediaFire's team upholds and enforces its policies and acts on the reported violations. MediaFire additionally employs a variety of processes and automatic mechanisms to avert violations of its ToS, which include: Media Fingerprinting Archive Scanning Monitoring websites

	Realtime Filters Blocking websites
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	When MediaFire removes or disables content due to policy violations, the user who posted the content may receive a strike. Repeated policy violations may result in account termination A confirmed report of a violation can result in actions up to and including immediate account termination.
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.

19. Google Docs

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.[206]), which apply to Google Docs, 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 marginalisation. *Terrorist activities*: Google does not permit terrorist organisations 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 organisations' 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://workspace.google.com/terms/premier_terms.html (Google Docs is part of Google Workspace) 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.[207]).
4.1 Notifications of removals or other enforcement decisions	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. If a user materially or repeatedly violates Google's Program Policies, Google may suspend or permanently disable that user's access to Google Docs. Google gives prior notice in such cases. However, Google may immediately suspend a user's use of Google Docs if Google believes that immediate suspension is required to comply with any applicable law

4.2 Appeal processes against removals or other enforcement decisions	If a file has a violation notice, the owner can request a review of the violation.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Users can report violating content. 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[208]). Google has reported that it monitors Google Docs based on an automated system of pattern matching that scans for indicators of violating content (Fung, 2017[233]).
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.[210]) - Report illegal materials to appropriate law enforcement authorities
7. Does the service issue transparency reports (TRs) on TVEC?	No. Google issues TRs (Google, n.d.[211]) encompassing Google's products and services, including Google Docs. 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.

20. Mega.nz

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, Mega's 2021 Transparency Report explains that Mega has zero tolerance for user sharing Objectionable material (as defined in section 3 of the New Zealand Films, Videos and Publications Classification Act 1993), which includes violent extremism (Mega.nz, 2021[62]).
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Mega's ToS are available at https://mega.io/terms Takedown Guidance Policy is available at https://mega.io/takedown
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?	Mega explains that it is impossible for Mega to review content uploaded by users, as it is encrypted on the user's device before it is sent to Mega. Nonetheless, even if the decryption key is provided to staff or otherwise publicly available, MEGA generally will not view, or attempt to view, files against which action is requested, but it reserves the right to do so where the file decryption key has been provided if it considers review is necessary or appropriate. MEGA is not obliged to take action unless required to do so by applicable law but any action will be undertaken objectively, based only on the information provided by third parties and its guidance and policies.
4.1 Notifications of removals or other enforcement decisions	MEGA will promptly inform the user of any action taken where practicable, provided it considers it appropriate or is required to do so by applicable law, and provided it is not legally prevented from doing so by a court or other authority with appropriate jurisdiction. However, action taken might not be disclosed in cases where an appropriate law enforcement agency requests non-disclosure because the case is under active investigation.
4.2 Appeal processes against removals or other enforcement decisions	Appeals against account closure for holding alleged objectionable material are referred to the New Zealand Authorities for adjudication of the content. The account can be reinstated if the content is determined to be not illegal.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Users can submit reports of links to objectionable material by email to abuse@mega.nz or by following a 'Report abuse' button on the page where a link is displayed.

	Most reports of violent extremism links are provided by an NGO which monitors public websites. Reports also come from law enforcement agencies, individuals, and industry actors. Mega does not scan stored files and then compare hash values to industry hash sets because user files are encrypted on the user's device before being uploaded and the stored encrypted file has a different hash to the original file. Mega is a member of the GIFCT.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Any reports of objectionable content result in immediate deactivation of the folder/file links, closure of the user's account, and provision of the details to the New Zealand Government Authorities for investigation and prosecution.
7. Does the service issue transparency reports (TRs) on TVEC?	Yes. Mega has published 7 TR since it commenced operations in January 2013.
8. What information/fields of data are included in the TRs?	In its last TR (which cover the year ended on 30 September 2021), Mega reported: - The total number of accounts closed to date for sharing objectionable content (which include violent extremism); - The number of reported violent extremism links that were disabled (numbers reported cover Q4 2019, 2020 and 2021); - The number of violent extremism link reports broken down by sources - NGOs, law enforcement, individuals, industry (numbers reported cover Q4 2019, 2020 and 2021) - The number of warrants Mega received during the reporting period for violent extremism, indicating the originating country and the outcome of the warrant (e.g. metadata supplied)
9. Methodologies for determining/ calculating/estimating the information/ data included in the TRs	No information provided.
10. Frequency/timing with which TRs are issued	On a yearly basis.

21. Pixeldrain

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No specific definition is provided. However, Pixeldrain's Content Policy prohibits content containing "terrorism", i.e. videos, images or audio fragments which promote and glorify acts of terrorism.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Pixeldrain's Content Policy is available at https://pixeldrain.com/about#content-policy
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?	No policies or 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 violating content using the report button on the download page of the file. When a file has received enough reports of the same type it will automatically be blocked. Staff moderators manually review reported files occasionally.
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.

22. Uploadgram

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There are neither ToS nor Community Guidelines/Standards. Thus, there is neither a specific definition of nor a general prohibition on TVEC.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	There are neither ToS nor Community Guidelines/Standards.
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?	Not applicable.
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)	No information available.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Not 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.

23. File.io

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No specific information is provided. However, File.io's ToS prohibit content that is hate speech, threatening or pornographic, that incites violence or that contains graphic or gratuitous violence.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	File.io's ToS are available at https://www.file.io/tos
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?	File.io broadly states that it may review content to determine whether it is illegal or violates its policies, and it may remove or refuse to display content that it believes violates its policies or the law. However, File.io does not generally review content beforehand, and it is not obligated to do so.
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 available.

6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Breach of File.io's ToS entitle File.io to suspend or stop the provision of its services to 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.

24. Gofile.io

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No specific information is provided. Gofile's ToS provide that users may not store, use, download, upload, share, access, transmit, or otherwise make available data in violation of any law in any country; abuse, defame, threaten, stalk or harass anyone, or harm them as defined by any law in any jurisdiction; and store, use, download, upload, share, access, transmit, or otherwise make available, unsuitable offensive, obscene or discriminatory information of any kind.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Gofile.io's ToS are available at https://gofile.io/terms
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?	Gofile.io broadly states that it reserves the right to remove data alleged to be infringing without prior notice, at its sole discretion.

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 available.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	According to Gofile.io, in appropriate circumstances, it will terminate a user's account if it considers that user to be a repeat 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.

25. Anonfiles

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No specific definition is provided. Anonfiles' ToS prohibit the distribution of illegal material via Anonfiles.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Anonfiles' ToS are available at https://anonfiles.com/terms
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	Anonfiles broadly states that its administrators have the right to remove and/or permanently ban file content they find inappropriate.

there notifications of removals or other enforcement decisions and appeal processes against them?	
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 violations of Anonfiles' ToS using a reporting form available at https://anonfiles.com/abuse.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Violations of Anonfiles' ToS entitle Anonfile to remove and/or permanently ban the violating content.
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.

Far-right-focused TVEC-intensive Services

26. Bitchute.com

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No specific definition is provided. However, in Bitchute.com's Community Guidelines, under the "Prohibited Content" heading, any individuals, organisations or other entities that are engaged in the following activities; any material that is produced by, endorsing, empowering or otherwise promoting individuals, organisations or any other entity engaging in the following activities; and posting, celebrating, endorsing, glorifying, denying the existence of, linking to or otherwise promoting content containing the following activities are not permitted

	to have a presence on Bitchute.com and are strictly prohibited: **Terrorism & Violent Extremism** Defined as any act of violence or intimidation carried out with the intention of furthering a religious, political or any other ideological objective. Entities that have been designated under counterterrorism legislation will be blocked within the jurisdiction of the relevant nation state or international organisation. In addition, those designated by the United Kingdom of Great Britain, Australia, Canada, New Zealand, the United States of America, the United Nations, the European Union or any member state of the European Union will be prohibited on the platform. BitChute maintains and publishes a Prohibited Entities List that contains entities that BitChute has independently identified and explicitly prohibited on the platform under this guideline. As this list will evolve over time, BitChute suggests that all users regularly check it to ensure they are not breaching the guidelines. The list can be found at https://support.bitchute.com/policy/prohibited-entities-list **Threats or Incitement to Violence** Defined as containing threats of violence or likely to incite violence. **Abhorrent Violence** Defined as real-life non-consensual acts of kidnapping, attempted murder, murder, mutilation, rape or torture. **Harmful Activities** Defined as the injection / ingestion of dangerous substances, self-harm, suicide and other activities that are intended to lead to someone getting badly hurt or worse. **Incitement to Hatred (UK, EU, EEA & territories)** As defined in Section 368E Subsection (1) of the UK Communications Act 2003, this applies to any material likely to incite hatred against a group of persons or a member of a group of persons based on any of the grounds referred to in Article 21 of the Charter of Fundamental Rights of the European Union (BitChute, 2021[234]).

2. Manner in which the ToS or Community Guidelines/Standards are communicated	BitChute's Community Guidelines are available at https://support.bitchute.com/policy/guidelines BitChute's Content Moderation policy is available at https://support.bitchute.com/policy/content-moderation
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?	Users can adjust their content's sensitivity level via their account settings. By default, all content is marked as Normal sensitivity and is visible to all users. Creators are expected to mark their content as Not Safe For Work (NSFW) or Not Safe For Life (NSFL) if appropriate before making the content public. Content Moderation processes are employed to ensure the appropriate sensitivity level is applied to content, and inappropriately marked content can be flagged via those processes when encountered. A creator's ability to change the sensitivity of a piece of content will be removed if the sensitivity has been set through the moderation process. Content sensitivity levels are defined as follows: **Normal** This is the default sensitivity level for content on the platform. Content should only be left as Normal sensitivity if it does not meet the thresholds to be marked Not Safe For Work (NSFW) or Not Safe For Life (NSFL). When viewing, Normal content should be considered equivalent to the British Board of Film Classification (BBFC) '12' rating. **Not Safe For Work (NSFW)** Content that is not safe for viewing in the workplace, or similar environments, must be marked as such by the content creator prior to publishing. Content containing discriminatory language, drug use, nudity and/or moderate violence should be marked as Not Safe For Work (NSFW). Not Safe For Work (NSFW) is equivalent to the British Board of Film Classification (BBFC) '15' rating. **Not Safe For Life (NSFL)**

	This level of sensitivity goes beyond Not Safe For Work (NSFW), as it does not matter where you view the material; many if not most people will find this content upsetting. Material that contains graphic content, which should be viewed with discretion, must be marked as Not Safe For Life (NSFL) by the content creator prior to publishing. Not Safe For Life (NSFL) is equivalent to the British Board of Film Classification (BBFC) '18' rating.
4.1 Notifications of removals or other enforcement decisions	No notifications are specified. However, given the manner in which appeals are made (see section 4.2 below), it seems that censored pages are labelled as such, and this labelling could be thus deemed a notification.
4.2 Appeal processes against removals or other enforcement decisions	In order to prevent platform bias, BitChute strives to make all moderation decisions made by the platform and its staff as transparent as legally possible, and subject to reasonable appeal. Users can only appeal content that they have posted to the platform, and they must include detailed reasoning for why it is believed the moderation action was incorrect. Appeals with no reasoning are automatically rejected. The recommended method for appealing content is manually via the Appeal function on the context navigation bar. Users must navigate to the page intended to be appealed, click on the Appeal icon, and then follow the instructions provided.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	When users discover prohibited material on BitChute, they can report it using the Flagging & Reporting tools provided. Reports can be made via the Bitchute.com platform, via email and via the Bitchute.com API. When requested to do so by the authorities, BitChute may apply filters on specific content that is considered illegal within their country. These filters will be applied so that they do not impact viewing of the content in other countries where the content is still considered legal.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Repeated offences under the Community Guidelines will lead to account suspension or termination.
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.

27. Rumble.com

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	Rumble does not specifically define "TVEC", but does prohibit certain categories of content, as more fully detailed under Rumble's current Content Policies, available at https://rumble.com/s/terms. Such prohibited content includes (i) content or material that is grossly offensive to the online community, including but not limited to, racism, antisemitism and hatred, (ii) content that promotes, supports, or incites violence or unlawful acts, (iii) content that promotes, supports or incites individuals and/or groups which engage in violence or unlawful acts, including but not limited to Antifa groups and persons affiliated with Antifa, the KKK and white supremacist groups and or persons affiliated with these groups, and (iv) content that promotes or supports entities and/or persons designated by either the Canadian or United States government as terrorists or terrorist organizations. Rumble recently issued a press release outlining a proposed new content policy and removal and appeal process (collectively the "Rumble Rules"). Rumble is seeking feedback from Rumble creators and users on its proposed Rumble Rules. The launch of the final Rumble Rules is expected by the end of this year. The press release and proposed new Rumble Rules are available at https://corp.rumble.com/blog/rumble-proposes-an-open-source-content-moderation-policy-process-to-improve-transparency-put-creators-first/.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Rumble's current Content Policies are available at https://rumble.com/s/terms. Information about Rumble's proposed future Rumble Rules is available at https://corp.rumble.com/blog/rumble-proposes-an-open-source-content-moderation-policy-process-to-improve-transparency-put-creators-first/.

3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	There is livestream functionality, but not specific provisions governing it.
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?	Rumble's functionality includes an online forum, live chat and comments for "Creator Discussions" (the "Forum") where users of the Rumble Service and creators of content may discuss matters pertaining to the content and/or the Rumble Service. Rumble reserves the right to monitor messages on comments and the Forum and to remove messages which Rumble in its sole discretion determines to be undesirable, inciting violence, harmful, offensive or otherwise in violation of its ToS. Any materials submitted to the Rumble Service may be, but are not necessarily, examined by Rumble before they are made available on Rumble. Rumble's proposed Rumble Rules contain a removal and appeal process. Under the current draft of the Rumble Rules, Rumble will notify a content creator of removal of "Contravening Content" (as defined in the Rumble Rules) if Rumble deems the "Community Identification" (as defined in the Rumble Rules) to be legitimate or otherwise substantiated and thus removes such Contravening Content.
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. However, the proposed Rumble rules do contemplate an appeal process.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	If a user has a complaint regarding content or materials available on Rumble, inappropriate behaviour or postings by other users in the Forum, or otherwise, they may submit the complaint to Rumble by emailing Rumble's customer service representatives at support@rumble.com. If the complaint concerns the activities of other users/visitors on the Forum, users must identify the specific type of inappropriate or offensive behaviour engaged in and, insofar as possible, the identity of the offending person. If the complaint concerns particular content, the reason for the complaint and the title and/or location of any video must be provided, as well as the date(s) on which the objectionable activities or behaviour were observed, or the

	date on which the content which is the subject of the complaint was viewed. Rumble observes that a customer service representative will endeavour to respond to the email, and if in Rumble's determination the complaint is a valid one, Rumble will take appropriate actions in its sole discretion, yet it has no responsibility at any time to report to the complainant as to the status or outcome of its investigation or any actions Rumble has taken as a result. Algorithms are not used to filter high risk video content; video content is subject to human review. According to Rumble's CEO Chris Pavloski, algorithms are mainly involved when "trying to figure out which videos are viral and which videos we need to put humans on to look at to distribute" (Kulvi, 2021[235]). Under the proposed Rumble Rules, Rumble will not use automated flagging for the identification of prohibited content save for copyright infringement and pornographic content. Rumble observes that its future content moderation policies will rely upon content creator and consumer flagging, as per the proposed Rumble Rules.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Rumble reserves the right, in its sole discretion, to terminate access to Rumble, with or without notice, for any reason, including, without limitation, if Rumble believes that a user has violated or acted inconsistently with the letter or spirit of Rumble's ToS. This includes Rumble's right to terminate the user's ability to upload videos, post comments, collect revenue or any function available via Rumble. Rumble has a zero tolerance for any violation of content polices and/or conduct outlined in its ToS. If a user is found in violation, their account may be suspended and/or terminated. The determination of suspension or termination is at the sole discretion of Rumble. Under Rumble's proposed future Rumble Rules, Rumble will maintain a right to immediately terminate an account if a content creator has received a removal notice for content that, in Rumble's opinion, is sufficiently egregious or clearly in violation of any applicable law. Rumble has also proposed a multi-level sanction and removal process for both substantiated and unsubstantiated complaints. All content removals will be subject to appeal.
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.

28. Gab.com

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No specific definition is provided. However, Gab's ToS provide that content and materials posted by users (User Contributions) must not: Be unlawful or be made in furtherance of any unlawful purpose. User Contributions must not aid, abet, assist, counsel, procure or solicit the commission of, nor constitute an attempt or part of a conspiracy to commit, any unlawful act. Gab notes for avoidance of doubt that speech which is merely offensive or the expression of an offensive or controversial idea or opinion, as a general rule, will be in poor taste but will not be illegal in the United States.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Gab's ToS are available at https://gab.com/about/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 enforcement decisions and appeal processes against them?	Gab strives to ensure that the First Amendment remains the Website's standard for content moderation. Gab makes best efforts to ensure that all content moderation decisions and enforcement of its ToS does not punish users for exercising their right to freedom of speech. According to Gab, it comparatively collects little data on users relative to other social networking sites. Gab's default position is that it should implement no prior restraints on any User Contribution. However, given the breadth of speech permitted on Gab, there may be circumstances where it is unable to determine whether content is protected by the First Amendment or not, in which cases Gab prefers to err on the side of caution. Accordingly, Gab reserves the right

	to take any action with respect to any User Contribution that it deems necessary or appropriate in its sole discretion, including the following: - Take any action with respect to any User Contribution that Gab deems necessary or appropriate in its sole discretion, including if Gab believe that such User Contribution violates its ToS, infringes any intellectual property right or other right of any person or entity, or could threaten the physical safety of users of the Website or the public. - Take appropriate legal action, including without limitation referral to law enforcement, for any illegal or unauthorized use of the Gab website or in cases of life-threatening emergency. - Terminate or suspend a user's access to all or part of the Gab Website for any violation of its 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. However, Gab notes that if a user's access to Gab is terminated or suspended in relation to a User Contribution that the user who authored it believes constitutes protected political or religious speech, and the user is able to demonstrate that the User Contribution in question was protected by the First Amendment by obtaining a declaratory judgment from a court of competent jurisdiction, Gab will consider permitting the user to re-join the site.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Gab observes that it does not review material before it is posted on the Gab, and cannot ensure prompt removal of unlawful material after it has been posted.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Gab reserves the right to disable any user name, password, or other identifier, whether chosen by a user or provided by Gab, at any time, if Gab believes the user has violated any provision of its ToS.
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.

29. Patriots.win

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No specific definition is provided.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Patriots.win's ToS are available at https://patriots.win/tos
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?	Neither policies nor 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 available.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Violation of patriots.win's ToS may lead to the termination of the infringer's account and viewing rights.

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.

30. Parler.com

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	- No specific definition is provided. However, Parler's Legal Guidelines provide that terrorist organizations officially recognized as such by the United States are forbidden from using Parler, as is anyone—including state actors—recruiting for them. - Moreover, reported "parleys" (i.e. posts), comments, or messages are deemed a violation of Parler's Legal Guidelines – and therefore are prohibited - if they contain: - a "serious expression of an intent to commit an act of unlawful violence to a particular individual or group of individuals," with either the intent or reckless disregard as to whether the communication will "place the victim in fear of bodily harm or death"; and - an explicit or implicit encouragement to use violence, or to commit a lawless action, such that: (a) the user intends his or her speech to result in the use of violence or lawless action, and (b) the use of violence or lawless action is the likely result of the parley, comment, or message.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Parler's ToS are available at https://parler.com/documents/termsofservice.pdf , the Legal Guidelines are available at https://legal.parler.com/documents/Elaboration-on-Guidelines.pdf , and the Community Guidelines are available at https://parler.com/documents/guidelines.pdf
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?	Parler states the following: "We do not curate your feed; we do not pretend to be qualified to do so. We believe only you are qualified to curate your feed, and so we give you the tools you need to do it yourself. To that end, Parler offers a number of features—including the ability to mute or block other users, or to mute or block all comments containing terms of your choice—and we encourage you to use these tools whenever the content you would rather not encounter here, is not otherwise addressed […]" (Parler, 2021[236])
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)	According to Parler: "We prefer that removing users or user-provided content be kept to the absolute minimum. We prefer to leave decisions about what is seen and who is heard to each individual. In no case will Parler decide what will content be removed or filtered, or whose account will be removed, on the basis of the opinion expressed within the content at issue. Parler's policies are, to use a well-known concept in First Amendment law, viewpoint-neutral. Parler will not knowingly allow itself to be used as a tool for crime, civil torts, or other unlawful acts. We will remove reported user content that a reasonable and objective observer would believe constitutes or evidences such activity. We may also remove the accounts of users who use our platform in this way" (Parler, 2021[236]). Parler has a reporting mechanism under which content deemed reproachable is brought to the attention of a "Community Jury", a group of user volunteers who vote on what violates Parler's rules. When the Community Jury votes that a given piece of content is prohibited, the content is removed, and actions may be taken against the infringer (see section 6 below). According to Parler:

	"Sometimes the law properly requires us to exclude content from our platform once it is reported to us or to our Community Jury—content we would make it a priority to exclude anyway. Obvious examples include: child sexual abuse material, content posted by or on behalf of terrorist organizations, intellectual property theft" (Parler, 2021[236]).
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Parler may remove any content and terminate a user's access to Parler at any time and for any reason to the extent that Parler reasonably believes (a) the user has violated Parler's ToS or Community Guidelines, (b) the user creates risk or possible legal exposure for Parler, or (c) the user is otherwise engaging in unlawful conduct.
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.

31. Odysee.com

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No specific definition is provided. However, Odysee's Community Guidelines prohibit any content or posts that incite hatred or violence towards a particular group or person(s) based on, but not limited to ethnicity, disability, nationality, race, gender, religion, sexual orientation, social class/caste, and gender identity/expression. Also, content or posts that promote terrorism, criminal activity, or credibly calls for violence (coordinated or otherwise), such as for example: sincere encouragement of others to go to a particular place to commit/perform violence, or to target groups or individuals with violence; promotion of recruitment into terrorist and/or criminal groups; sincere promotion of terrorist and/or criminal groups; and sincere promotion of terrorism and/or criminal activity are also prohibited. When content related to terrorism or crime is posted for an educational, documentary, scientific, or artistic purpose,

	users must be mindful to provide enough information in the video or audio itself so viewers understand the context.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Odysee's Community Guidelines are available at https://odysee.com/@OdyseeHelp:b/Community-Guidelines:c , ToS are available at https://odysee.com/$/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 enforcement decisions and appeal processes against them?	Odysee observes that that there is no such thing as a one size fits all approach to moderation, so it encourages its users to take advantage of additional moderation tools available to them and shape their experience on Odysee in a way that aligns with their personal values. In particular, channel creators can enable and disable comments, switch to "slow mode" (which limits how quickly users can leave new chats/comments) and block users. Creators can delegate other users as moderators. Moderators have the same ability to block users and remove comments as the creator. Creators and moderators also have the ability to remove any content posted in the relevant channel (Odysee, 2021[237]).
4.1 Notifications of removals or other enforcement decisions	No notifications are specified.
4.2 Appeal processes against removals or other enforcement decisions	Odysee notes that if a user believes a moderation decision was wrong, the user is "welcome to submit feedback" to Odysee via hello@odysee.com. The user must reference the relevant video url in the feedback form.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	Odysee has a reporting tool allowing users to report violating content. Odysee's moderators review the reports and take action when the violation is confirmed (see section 6 below).

6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	When violating content is found, Odysee request for immediate removal. Odysee moves forward with removal where the creator or user is unable to. In circumstances where there are repeated breaches of Odysee's community guidelines, Odysee may pursue more stricter action(s). For example: - Filtering of the infringer's channel from Odysee; or - Restricting the infringer's ability to comment, either temporarily or permanently.
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.

32. Brandnewtube.com

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No specific definition is provided. Brandnewtube broadly prohibits use of its website in a manner inconsistent with any applicable laws or regulations.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	ToS are available at https://brandnewtube.com/terms/terms
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	There is livestream functionality, yet no specific provisions governing it.
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 information provided.

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 provided.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Violation of Brandnewtube's ToS may lead to the suspension or termination of the infringer's account, in which case the infringer is denied any and all current or future use of Brandnewtube.com.
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.

33. Gettr.com

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No specific definition is provided. However, under Gettr's ToS, users may not post on or transmit any unlawful, harmful, threatening, abusive, harassing, defamatory, libelous, indecent, vulgar, obscene, sexually explicit, pornographic, profane, hateful, racially, ethnically or otherwise objectionable material of any kind, including any material that encourages conduct that would constitute a criminal offense, give rise to civil liability or otherwise violate any law, rule or regulation of the laws applicable to users or applicable in the country in which the material is posted. For example, this may include content identified as personal bullying, sexual abuse of a child, attacking any religion or race, or content containing video or depictions of beheadings.

2. Manner in which the ToS or Community Guidelines/Standards are communicated	Gettr's ToS are available at https://gettr.com/terms
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	There is livestream functionality, yet no specific provisions governing it.
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?	Gettr observes that it may, but will not have any obligation to, review, monitor, display, post, store, maintain, accept, or otherwise make use of, any of a user's user-generated content (UGC), and Gettr may, in its sole discretion, reject, delete, move, re-format, remove, or refuse to post or otherwise make use of UGC without notice or any liability to the user or any third-party in connection with Gettr's operation of UGC venues in an appropriate manner, such as to enhance accessibility of UGC, address copyright infringement, and protect users from harmful UGC. Without limitation, Gettr may, but does not commit to, do so to address content that comes to its attention that it believes is offensive, obscene, lewd, lascivious, filthy, pornographic, violent, harassing, threatening, abusive, illegal, or otherwise objectionable or inappropriate, or to enforce the rights of third parties or its ToS. For example, this may include content identified as personal bullying, sexual abuse of a child, attacking any religion or race, or content containing video or depictions of beheadings.
4.1 Notifications of removals or other enforcement decisions	In certain cases, in Gettr's sole discretion, it may provide a user with a written notice (a "Restriction Notice") to inform that: (i) the user's right to use or access any part of Gettr has been terminated, including the right to use, access or create any account thereon; and (ii) Gettr refuses to provide any service to such user.
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 violating content by contacting Gettr. Nonetheless, Gettr states that it assumes no responsibility for ongoing monitoring of the Interactive Community (basically any communication functionality on the website) or for removal or editing of any UGC, even after receiving notice.

6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Gettr may, upon notice to the relevant user, issue a warning, temporarily suspend, indefinitely suspend, or terminate the user's account or access to all or any part of Gettr for any reason, in its sole discretion, including for violations of its ToS.
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.

34. 8kun.top

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There are neither ToS nor Community Guidelines/Standards. Thus, there is neither a specific definition of nor a general prohibition on TVEC. However, in 8kun's homepage, it is stated that any content that violates the laws of the United States of America will be deleted and the poster will be banned.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	There are no ToS. However, there is a privacy policy with information on 8kun's operational aspects, which is available at https://8kun.top/privacy.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?	There are neither policies, procedures nor ToS. Nonetheless, 8kun's privacy policy explains that accounts are not needed to use 8kun unless a user wishes to create and own a board and moderate it. In this case, as a Board Owner, the user may hire Board Volunteers and/or Reporters for his or her board, but this is not a requirement. Board Volunteers help with moderation duties (e.g. ban users or delete content). Reporters are designated to create

	threads, but this is not a requirement. Reporters are not able to ban, delete or otherwise moderate boards. 8kun allows users to upload, publish, and share text and media ("Posts"). 8kun does not exercise any control over Posts or any personal information embodied therein other than receiving and storing Posts at the direction of users and, in some circumstances (which are not specified), deleting and/or banning Posts. When a user publishes a Post on 8kun, his or her IP address is hashed with a private salt, which is rotated periodically, for privacy-preserving purposes. 8kun stores the hashed IP addresses and other post parameters, such as the message body and the Subject, Email, and Name fields, if they were filled out. Any image/s or video/s uploaded by users with their Posts, if applicable, is/are also stored. This information is retained until one of the following happens: - From a local standpoint, the Post is deleted by the Board Owner or one of his or her volunteers; - From a global standpoint, the Post is deleted by a Global Volunteer, an Operator, or any other member of the 8kun Administration; - From the user's standpoint, the Post is deleted by the user him/herself if the Board Owner has post deletion enabled on his or her board. The Post data is pruned automatically, for instance, when a board's catalog has reached the maximum number of threads allowed. When a hashed IP address is banned, it is added to two separate tables: bans and user action log. All hashed IP addresses of active users are checked against the bans table to make sure users who were banned, for whatever reason, are unable to post until their bans have expired. The user action log table is used for generating all board logs across 8kun. So, if a hashed IP address is banned, even for a second, it will remain stored in the user action log table forever, even though it might have already left the bans table after the ban expired. As for permanent global or local bans, the hashed IP address will remain stored in both tables forever, unless bans are reset because of a salt rotation or the hashed IP address is manually unbanned for whatever reason, in

	which case the hashed IP address will leave the bans table, but remain in the user action log table.
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)	The explanation summarised in section 4 above (taken from 8kun's privacy policy) suggests that global (website level as opposed to board level) volunteer moderators and staff moderators have the ability to delete posts, but it is unknown under what circumstances and on what grounds.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Not 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.

35. Redvoicemedia.com

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There are neither ToS nor Community Guidelines/Standards. Thus, there is neither a specific definition of nor a general prohibition on TVEC.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	There are neither ToS nor Community Guidelines/Standards (only a privacy policy with very limited information).
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?	Not applicable.
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 available.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Not 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.

36. Thedonald.win

As of 3 March 2022, this website is not operational. Upon entering thedonald.win on a web browser, users are redirected to www.america.win. On this website there is a statement posted by Jody Williams, a Trump supporter who explains what events motivated the demise of thedonald.win (which seems to have been decommissioned between February-March 2021).

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	Not applicable.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Not applicable.
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?	Not applicable.
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)	Not applicable.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Not applicable.
7. Does the service issue transparency reports (TRs) on TVEC?	Not applicable.
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.

37. WeGo.Social

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No specific definition is provided. WeGo.Social's ToS prohibit use of WeGo.Social in a manner inconsistent with any applicable laws or regulations.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	WeGo.Social's ToS are available at https://wego.social/terms/terms
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?	WeGo.Social broadly states that it has the right to remove any user content put on the website if, in its opinion, such user content does not comply with WeGo.Social's 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)	Users can complain about user content uploaded by other users at admin@wg1wga.com. On the section "about WeGo.social" of WeGo.Social's website, it is stated that WeGo.Social is "community moderated" (WeGo.Social, n.d.[238]). This type of moderation is explained as follows: "The days of trolls ruining the social media experience has come to an end. Our TG1TGA feature, short for TheyGo1TheyGoAll, is our ultimate tool to fight this. If a troll decides not to act accordingly our community of patriot users can give them a TG1TGA with the click of their mouse. Get too many TG1TGA's and the troll is suspended

	from the site for a period of time while also losing points, giving the power to you, the people. Patriot admins on the backend monitor how TG1TGA's are used to prevent abuse but remember numbers matter so no single person can make the decision to suspend a user" (WeGo.Social, n.d.[238]). The passage above suggests that all users are moderators based on a voting mechanism, and that staff moderators monitor potential abuses of this system.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	If WeGo.Social determines that a user's use of WeGo.Social is in breach of its ToS or of any applicable law or regulation, WeGo.Social may terminate that user's use or participation on WeGo.Social or delete their profile and any content or information they posted at any time, without warning, in WeGo.Social'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.

38. Safechat.com

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No specific definition is provided. However, SafeChat's ToS prohibit any content that is abusive, hateful, threatening, and harmful, incites violence; or contains graphic or gratuitous violence.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	SafeChat's ToS are available at https://help.safechat.com/terms/ . Although SafeChat's ToS made reference to some "Community Standards", these are nowhere to be found.
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?	SafeChat's ToS highlight its right to monitor activity and content associated with SafeChat, but also provide that Safechat is not obligated to do so. Safechat reserves the right to remove objectionable content without notice. Also, SafeChat removes any content or information users post on SafeChat if it believes that it violates its ToS.
4.1 Notifications of removals or other enforcement decisions	SafeChat reserves the right to ban Channels, Groups, and users that do not comply with its Terms of Service as assessed in SafeChat's sole discretion and interpretation of its ToS; and Users, Channels, and Groups under investigation or which have been detected as sharing content in violation of such ToS may have their visibility limited in various parts of Safecha, including search. Channels, Groups, and users may not be notified when any of this occurs. This suggests that notifications are at SafeChat's discretion.
4.2 Appeal processes against removals or other enforcement decisions	Banned Channels, Groups, and users may contact SafeChat Customer Support to request more information about a moderation decision. SafeChat may or may not respond to such requests.
5. Means of identifying TVEC (for example, monitoring algorithms, user generated, human (staff) reviewers, hash-sharing/URL sharing database)	SafeChat observes that it may investigate the complaints and violations of its policies that come to its attention and may take any action that it believes is appropriate, including, but not limited to issuing warnings, removing the content, or terminating accounts and/or subscriptions. However, SafeChat reserves the right not to take any action.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	SafeChat may suspend or terminate a user's account or cease providing them with all or part of SafeChat's services at any time for any or no reason, including if SafeChat reasonably believes the user has violated its ToS (see also sections 4.1, 4.2 and 5 above).
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.

39. 88nsm.com

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There are neither ToS nor Community Guidelines/Standards. Thus, there is neither a specific definition of nor a general prohibition on TVEC.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	There are neither ToS nor Community Guidelines/Standards.
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?	Not applicable.
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)	No information available.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Not 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.

40. Doxbin.org

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No specific definition is provided. Doxbin's ToS clarify that there are no limitations on what info can be posted as long as it is not spam, child explicit material or violating hosting country jurisdictional laws. "For example, kiddie porn links, your mothers creditcard and any support of terrorism is not allowed here" (Doxbin, n.d.[239]).
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Doxbin's ToS are available at https://doxbin.com/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 enforcement decisions and appeal processes against them?	Doxbin broadly states that any pastes (i.e. posts) directly threatening and/or attempting to injure or hurt any particular individuals will be swiftly removed.
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 wishing to request for a paste to be removed must contact Doxbin's staff on Telegram at @Brenton or @Doxer.

6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Pastes that break Doxbin's ToS are subject to be removed.
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.

41. Wimkin.com

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No specific definition is provided. However, Wimkin's ToS provide that users' posts (Contributions) must not be threatening nor criminally harassing in any way shape or form. Criminal intent including posts calling for the organization of violence are also prohibited.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Wimkin's ToS are available at https://wimkin.com/terms/
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?	Wimkin observes that it monitors the website "24 hours per day, 7 days per week". Wimkin tries to check that all submissions comply with its acceptable use standards (contained in its ToS) "as soon as reasonably practicable after publication."
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)	Wimkin has a contact form to report "something that requires immediate attention". Moderators review and take action on these reports (Wimkin, 2021[240]). In particular, when a complaint is submitted, users must (1) outline the reason for the complaint, and (2) specify where the Contribution being complained about is located. Wimkin may request further information before processing the complaint. Then, Wimkin reviews the Contribution and decides whether it complies with its ToS. Wimkin informs the complainant of the outcome of its review within a reasonable time of receiving the complaint.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	If in Wimkin's opinion an individual makes use of the website in breach of its ToS, Wimkin may remove, or disable access to, any Contribution, and terminate, suspend or change the conditions of their access to the website without prior warning. Wimkin may also bring legal proceedings against any individual for a breach of its ToS, or take such other action as it reasonably deems appropriate. No nudity, pornography or criminal intent including posts calling for the organization of violence are permitted in Wimkin, and will result in immediate account removal and a ban on 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.

42. Mzwnews.com

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There are neither ToS nor Community Guidelines/Standards. Thus, there is neither a specific definition of nor a general prohibition on TVEC.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	There are neither ToS nor Community Guidelines/Standards.
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?	Not applicable.
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)	No information available.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Not 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.

43. Worldtruthvideos.website

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No specific definition is provided. Worldtruthvideos's ToS provide that "All legal content is allowed, any illegal content will be removed and your account may be terminated."
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Worldtruthvideos's ToS are available at https://worldtruthvideos.website/terms/terms
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 Worldtruthvideos, users can host videos "and know that they won't be taken down or censored" (Worldtruthvideos.website, 2022[241]).
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 available.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	According to Worldtruthvideos's ToS, "All legal content is allowed, any illegal content will be removed and your account may be terminated."

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.

44. Xephula.com

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No specific definition is provided. Xephula's ToS prohibit the use of XEPHULA in violation of any laws in users' jurisdiction, the USA and the state of Texas.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Xephula's ToS are available at https://xephula.com/static/terms
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?	Xephula broadly states that it may, but have no obligation to, remove content and accounts containing content that it determines in its sole discretion are unlawful or violate its 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)	Xephula has a contact form to report violating content.

6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Violation of any of Xephula's ToS may result in the 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.

45. Thegreaterreset.org

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There are neither ToS nor Community Guidelines/Standards. Thus, there is neither a specific definition of nor a general prohibition on TVEC.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	There are neither ToS nor Community Guidelines/Standards.
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?	Not applicable.
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)	No information available.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Not 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.

46. Nordfront.dk

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No specific definition is provided. Nordfront's Comment Rules provide that comments may not be discriminatory; incite, argue for, or contribute to violence or other illegal acts; express violent intentions towards people, organizations, buildings or the like; or be too violent / offensive.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Nordfront's Comment Rules are available at https://www.nordfront.dk/regler/
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?	Nordfront notes that its comment field is moderated, and if a breach of its rules is discovered, the comment will be removed or edited without further notice. Regardless of whether Nordfront receives an external complaint, or discover the breach itself, Nordfront decides on its own whether there is a breach of its rules. If so, Nordfront responds "as it sees fit."

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)	Comments that violate the Comments Rules can be reported via: info@nordfront.net, and staff moderators investigate the matter.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Repeated attempts to publish material that violates the Comment Rules may result in the blocking of the user / IP address.
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.

47. Lookaheadamerica.org

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There are neither ToS nor Community Guidelines/Standards. Thus, there is neither a specific definition of nor a general prohibition on TVEC.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	There are neither ToS nor Community Guidelines/Standards.
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	Not applicable.

there notifications of removals or other enforcement decisions and appeal processes against them?	
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)	No information available.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Not 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.

48. Patriotfront.us

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There are neither ToS nor Community Guidelines/Standards. Thus, there is neither a specific definition of nor a general prohibition on TVEC.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	There are neither ToS nor Community Guidelines/Standards.
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?	
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)	No information available.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Not 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.

49. Vastarinta.com

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There are neither ToS nor Community Guidelines/Standards. Thus, there is neither a specific definition of nor a general prohibition on TVEC.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	There are neither ToS nor Community Guidelines/Standards.

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?	Not applicable.
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)	No information available.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Not 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.

50. Nordicresistancemovement.org

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There are neither ToS nor Community Guidelines/Standards. Thus, there is neither a specific definition of nor a general prohibition on TVEC.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	There are neither ToS nor Community Guidelines/Standards.
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?	Not applicable.
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)	No information available.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Not 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.

Annex E - Definitions

For purposes of this report, the following definitions are provided:

Content: Any type of digital information serving as a medium for TVEC, such as comments, pictures, videos, files, posts, links, chatroom chats, blogs or messages.

Content-Sharing Service: Any online service that enables the transfer, transmission and dissemination of Content, in whatever form, whether one-to-one, one-to-few or one-to-many and irrespective of whether the Content is public-facing, semi-private or private. All of the Services profiled in this Report are Online Content-Sharing Services.

Online Platform: A digital service that facilitates interactions between two or more distinct but interdependent sets of users (whether firms or individuals) who interact through the service via the Internet.

Social Media (or Social Networking) Service: Any online service that allows individuals to build a public or semi-public profile of themselves, upload and access Content shared by other users, interact and establish connections with other users, and express their views and interests.

Terrorist Use of the Internet (TUI): Use of the Internet to promote terrorist aims (for example, using a messaging app to coordinate a terrorist attack). The dissemination of TVEC is a type of TUI whose purpose may be, for instance, to incite violence, radicalise or recruit.

Terrorist and Violent Extremist Content (TVEC): There is no universally accepted definition of terrorism and violent extremism, and congruently, of TVEC. This Report follows the language employed in the Christchurch Call, and uses these terms to refer to the general category of terrorist and violent extremist content on which several Online Content-Sharing Services have policies, make moderation and removal decisions, and in some cases report on in transparency reports.

References

99 Firms (2021), *LinkedIn Statistics*, https://99firms.com/blog/linkedin-statistics/. [81]

99 Firms (2021), *Viber Statistics*, https://99firms.com/blog/viber-statistics/. [76]

Access Now (2020), *Two Years under the EU GDPR*. [60]

Ahmed, M. (2020), *After Christchurch: How Policymakers Can Respond to Online Extremism*, https://institute.global/policy/after-christchurch-how-policymakers-can-respond-online-extremism. [271]

Alexander, J. (2019), *Verizon is selling Tumblr to WordPress' owner*, https://www.theverge.com/2019/8/12/20802639/tumblr-verizon-sold-wordpress-blogging-yahoo-adult-content. [164]

Anti-defamation League (2021), *"The Great Replacement:" An Explainer*, https://www.adl.org/resources/backgrounders/the-great-replacement-an-explainer. [38]

Apple (n.d.), *Privacy - About Apple's Transparency Report*, https://www.apple.com/legal/transparency/about.html. [131]

Article 19 (2020), *EU: Terrorist Content Regulation must protect freedom of expression rights*, https://www.article19.org/resources/eu-terrorist-content-regulation-must-protect-freedom-of-expression-rights/. [18]

Ask.fm (2021), *ASKfm Safety Guide for Schools & Educators*, https://safety.ask.fm/ask-fm-safety-guide-for-schools-educators/. [86]

Ask.fm (2021), *Transparency Report*, https://about.ask.fm/legal/en/transparency.html. [44]

Audiens (2020), *Game developer, Smule, increase their high value VIP customers by 21%*, https://audiens.com/smule-increase-their-high-value-vip-customers-by-21/. [99]

Automattic (n.d.), *Transparency Report*, https://transparency.automattic.com/iru-reports/. [221]

Barnes, L. (2019), *One month after controversial adult-content purge, far-right pages are thriving on Tumblr*, https://thinkprogress.org/far-right-content-survived-tumblr-purge-36635e6aba4b/. [167]

Barret, P. (2020), *Regulating Social Media: The Fight overn Section 230 - and Beyond*, NYU / STERN - Center for Business and Human Rights. [270]

BBC (2017), *WhatsApp must not be 'place for terrorists to hide'*, https://www.bbc.co.uk/news/uk-39396578. [24]

Bennett, C. (2019), *Extremism, George Washington University*, https://extremism.gwu.edu/sites/g/files/zaxdzs2191/f/EncryptedExtremism.pdf. [151]

Berger, J. and J. Morgan (2015), *The ISIS Twitter Census: Defining and Describing the Population of ISIS Supporters on Twitter*, The Brookings Institution. [20]

BitChute (2021), *Community Guidelines*, https://support.bitchute.com/policy/guidelines. [234]

Brandt, L. and G. Dean (2021), *Gab, a social-networking site popular among the far right, seems to be capitalising on Twitter bans and Parler being forced offline. It says it's gaining 10,000 new users an hour*, https://www.businessinsider.com.au/gab-reports-growth-in-the-midst-of-twitter-bans-2021-1. [26]

Broadcasting and Telecommunications Legislative Review Panel (2020), *Canada's Communications Future: Time to Act*. [310]

BSR (2021), *Human Rights Assessment: Global Internet Forum To Counter Terrorism*. [36]

Cambridge Consultants (2019), *Use of AI in Online Content Moderation*. [51]

Carmen, A. (2015), *Filtered extremism: how ISIS supporters use Instagram*, https://www.theverge.com/2015/12/9/9879308/isis-instagram-islamic-state-social-media. [135]

Chen, W. (2020), *The top Chinese short-video apps in 2020 vying to grab your attention with fast content*, https://kr-asia.com/the-top-chinese-short-video-apps-in-2020-vying-to-grab-your-attention-with-fast-content. [269]

Christchurch Call (2019), *Christchurch Call*, https://www.christchurchcall.com/call.html. [12]

Christmann, K. (2012), *Preventing Religious Radicalisation and Violent Extremism - A Systematic Review of the Research Evidence*. [268]

Clicky, S. (2017), *Tackling Extremist Content on WordPress.com*, https://transparency.automattic.com/2017/12/06/tackling-extremist-content-on-wordpress-com/. [219]

Clifford, B. and H. Powell (2019), *Encrypted Extremism - Inside the English-Speaking Islamic State Ecosystem on Telegram*. [267]

Commission for Countering Extremism (2019), *Challenging Hateful Extremism*. [309]

Conway, M. et al. (2019), "Disrupting Daesh: Measuring Takedown of Online Terrorist Material and Its Impacts", *Studies in Conflict & Terrorism*, Vol. 42/1-2. [21]

Counter Extremism Project (2018), *On Anniversary Of Barcelona Attacks, ISIS Continues Its Expansion*, https://www.counterextremism.com/press/anniversary-barcelona-attacks-isis-continues-its-expansion. [216]

Counter Terrorism Project (n.d.), *Extremists & Online Propaganda*, https://www.counterextremism.com/extremists-online-propaganda. [189]

Cox, J. (2019), *36 Days After Christchurch, Terrorist Attack Videos Are Still on Facebook*, https://www.vice.com/en_us/article/43jdbj/christchurch-attack-videos-still-on-facebook-instagram. [137]

Creemers, R. (2018), *newamerica.org*, https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-cybersecurity-law-peoples-republic-china/. [54]

Cuthbertson, A. (2019), *TikTok secretly loaded with Chinese surveillance software, lawsuit claims*, https://www.independent.co.uk/life-style/gadgets-and-tech/news/tiktok-china-data-privacy-lawsuit-bytedance-a9230426.html. [266]

Datanyze (2021), *File Sharing Software Market Share*, https://www.datanyze.com/market-share/file-sharing--198. [101]

Datareportal (2021), *Global Social Media Stats*, https://datareportal.com/social-media-users. [73]

Daum Kakao (2020), *Digital Responsibility for Protecting User Privacy: Kakao Transparency Report*, https://privacy.kakao.com/transparency/report?lang=en. [199]

DCMS (2020), *Online Harms White Paper - Initial consultation response*, https://www.gov.uk/government/consultations/online-harms-white-paper/public-feedback/online-harms-white-paper-initial-consultation-response. [297]

Dean, B. (2021), *Twitch Usage and Growth Statistics: How Many People Use Twitch in 2021?*, https://backlinko.com/twitch-users#monthly-active-users. [92]

Dearden, L. (2019), *Far-right extremists 'encouraged copycat terror attacks' after Christchurch mosque shootings*, https://www.independent.co.uk/news/uk/crime/far-right-terror-plots-uk-muslims-christchurch-attack-white-a9050511.html. [130]

Dempsey, F. (ed.) (2017), *Systematic Government Access to Private-Sector Data in China*, Oxford University Press. [242]

Department of Justice Canada (2021), *Government of Canada takes action to protect Canadians against hate speech and hate crimes*, https://www.canada.ca/en/department-justice/news/2021/06/government-of-canada-takes-action-to-protect-canadians-against-hate-speech-and-hate-crimes.html. [303]

DeviantArt (2021), *About DeviantArt*, https://www.deviantart.com/about/. [100]

DeviantArt (2020), *Proactive Deviation Moderation*, https://www.deviantart.com/team/journal/Proactive-Deviation-Moderation-848744195. [204]

DeviantArt (n.d.), *What happens when my account is banned?*, https://www.deviantartsupport.com/en/article/what-happens-when-my-account-is-banned. [203]

DeviantArt (n.d.), *What is your policy around account suspensions?*, https://www.deviantartsupport.com/en/article/what-is-your-policy-around-account-suspensions. [202]

DeviantArt (n.d.), *What policy guidelines are there on comments, Journals, statuses, and general interactions?*, https://www.deviantartsupport.com/en/article/what-policy-guidelines-are-there-on-comments-journals-statuses-and-general-interactions. [201]

Dilger, D. (2015), *Another security manual recommends using Apple iMessage: this time, ISIS*, https://appleinsider.com/articles/15/11/21/another-security-manual-recommends-using-apple-imessage-this-time-isis-. [133]

Discord (2021), *An Update on Our Business*, https://discord.com/blog/an-update-on-our-business. [91]

Discord (2021), *Announcing the Discord Moderator Academy Exam*, https://blog.discord.com/announcing-the-discord-moderator-academy-exam-a1bcb5b9d405. [186]

Discord (2021), *Discord Transparency Report: January - June 2021*, https://discord.com/blog/discord-transparency-report-h1-2021. [187]

Discord (2021), *Discord Transparency Report: July - December 2020*, https://discord.com/blog/discord-transparency-report-july-dec-2020. [188]

Discord (2021), *How Trust & Safety Addresses Violent Extremism on Discord*, https://discord.com/blog/how-trust-safety-addresses-violent-extremism-on-discord. [35]

Donovan, J., B. Lewis and B. Friedberg (2019), *arallel Ports: Sociotechnical Change from the Alt-Right to Alt-Tech*, Bielefeld. [28]

Doxbin (n.d.), *TOS*, https://doxbin.com/tos. [239]

Droogan, J., L. Waldek and R. Blackhall (2018), "Innovation and terror: an analysis of the use of social media by terror-related groups in the Asia Pacific", *Journal of Policing, Intelligence and Counter Terrorism*, Vol. 13/2. [3]

Dropbox (n.d.), *Transparency Overview*, https://www.dropbox.com/en_GB/transparency. [215]

Dropbox (n.d.), *Who can see the stuff in my Dropbox account? Dropbox Help*, https://help.dropbox.com/accounts-billing/security/file-access. [214]

Duarte, N., E. Llanso and A. Loup (2017), *Mixed Messages? The Limits of Automated Social Media Cotent Analysis*. [265]

Dusaleev, V. (2020), *Odnoklassniki: Robbie the Platform for the analysis of content*, https://tadviser.com/index.php/Product:Odnoklassniki:_Robbie_the_Platform_for_the_analysis_of_content. [194]

EDRi (2019), *Trilogues on terrorist content: Upload or re-upload filters? Eachy peachy*, https://edri.org/our-work/trilogues-on-terrorist-content-upload-or-re-upload-filters-eachy-peachy/. [275]

Electronic Frontier Foundation (2020), *Urgent: EARN IT Act Introduced in House of Representatives*, https://www.eff.org/deeplinks/2020/10/urgent-earn-it-act-introduced-house-representatives. [274]

Envisage Digital (2021), *WordPress Market Share in 2021*, https://www.envisagedigital.co.uk/wordpress-market-share/. [102]

eSafety Commissioner (2022), *Tech platforms asked to explain how they are tackling online child sexual exploitation*, https://www.esafety.gov.au/newsroom/media-releases/tech-platforms-asked-explain-how-they-are-tackling-online-child-sexual-exploitation. [68]

eSafety Commissioner (2021), *Development of industry codes under the Online Safety Act*, https://www.esafety.gov.au/sites/default/files/2021-09/eSafety%20Industry%20Codes%20Position%20Paper.pdf. [67]

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*. [294]

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*. [291]

European Commission (2020), *The Digital Services Act package*, https://ec.europa.eu/digital-single-market/en/digital-services-act-package. [287]

European Commission (2018), *Commission Reccomendation EU) 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online*. [8]

European Commission (2018), *Proposal for a Regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online - A contribution from the European Commission to the Leaders' meeting in Salzburg on 19-20 September 2018*. [2]

EUROPOL (2020), *European Union Terrorism Situation and Trend Report*. [6]

Facebook (2021), *Appealed Content*, https://transparency.fb.com/en-gb/policies/improving/appealed-content-metric/. [105]

Facebook (2021), *Content actioned*, https://transparency.fb.com/en-gb/policies/improving/content-actioned-metric/. [112]

Facebook (2021), *Disabling Accounts*, https://transparency.fb.com/en-gb/enforcement/taking-action/disabling-accounts/. [110]

Facebook (2021), *How Facebook enforces its policies*, https://transparency.fb.com/en-gb/enforcement/. [47]

Facebook (2021), *How review teams work*, https://transparency.fb.com/en-gb/enforcement/detecting-violations/how-review-teams-work/. [108]

Facebook (2021), *How Technology Detects Violations*, https://transparency.fb.com/en-gb/enforcement/detecting-violations/technology-detects-violations/. [107]

Facebook (2021), *Prevalence*, https://transparency.fb.com/en-gb/policies/improving/prevalence-metric/. [48]

Facebook (2021), *Taking down violating content*, https://transparency.fb.com/en-gb/enforcement/taking-action/taking-down-violating-content/. [109]

Facebook (2021), *Violence and Incitement*, https://transparency.fb.com/en-gb/policies/community-standards/violence-incitement/. [43]

Facebook (2020), *An Update on Combating Hate and Dangerous Organizations*, https://about.fb.com/news/2020/05/combating-hate-and-dangerous-organizations/. [312]

Facebook (2020), *An Update to How We Address Movements and Organizations Tied to Violence*, https://about.fb.com/news/2020/08/addressing-movements-and-organizations-tied-to-violence/. [34]

Facebook (2019), *Combating Hate and Extremism*, https://about.fb.com/news/2019/09/combating-hate-and-extremism/. [308]

Facebook (2017-2021), *Community Standards Enforcement Report*, https://transparency.fb.com/data/community-standards-enforcement/?from=https%3A%2F%2Ftransparency.facebook.com%2Fcommunity-standards-enforcement%2Fguide. [111]

Facebook (n.d.), *Community Standards, 1. Violence and Incitement*, https://www.facebook.com/communitystandards/credible_violence. [104]

Facebook (n.d.), *Community Standards, Dangerous Individuals and Organisations*, https://www.facebook.com/communitystandards/dangerous_individuals_organizations. [33]

Finances Online (2021), *Number of Tumblr Blogs in 2021/2022: User Demographics, Growth, and Revenue*, https://financesonline.com/number-of-tumblr-blogs/. [80]

Fisher-Birch, J. (2018), *Terror on Tumblr*, https://www.counterextremism.com/blog/terror-tumblr. [168]

Flickr (2021), *The Flickr Moderation Bot has changed the safety level of my photo*, https://www.flickrhelp.com/hc/en-us/articles/4405279492500-The-Flickr-Moderation-Bot-has-changed-the-safety-level-of-my-photo. [196]

Flickr (2021), *Work at Flickr*, https://www.flickr.com/jobs/. [97]

Frampton, M., A. Fisher and N. Prucha (2017), *The New Netwar: Countering Extremism Online*. [264]

Frier, S. (2018), *Facebook Scans the Photos and Links You Send on Messenger*, https://www.bloomberg.com/news/articles/2018-04-04/facebook-scans-what-you-send-to-other-people-on-messenger-app. [138]

Fung, B. (2017), *Why Google 'reads' your Docs*, https://www.stuff.co.nz/technology/digital-living/98470029/why-google-reads-your-docs. [233]

G20 (2019), *G20 Osaka Leaders' Statement on Preventing Exploitation of the Internet for Terrorism and Violent Extremism Conducive to Terrorism (VECT)*, https://dig.watch/instruments/g20-osaka-leaders-statement-preventing-exploitation-internet-terrorism-and-violent. [9]

G20 (2017), *The Hamburg G20 Leaders' Statement on Countering Terrorism*, https://www.mofa.go.jp/files/000271330.pdf. [11]

G7 (2019), *G7 Digital Ministers Chair's Summary*, https://www.economie.gouv.fr/files/files/2019/G7/G7Num/Chairs_summary_version_finale_ENG.pdf. [10]

GIFCT (2021), *HRIA Response Letter by Nick Rasmussen*, https://gifct.org/2021/07/20/hria-response-letter-by-nick-rasmussen/. [301]

GIFCT (2021), *Membership*, https://gifct.org/membership/. [298]

GIFCT (2021), *Membership*, https://gifct.org/membership/. [299]

GIFCT (2021), *Transparency Report*. [280]

GIFCT (2020), *GIFCT Transparency Report - July 2020*, https://gifct.org/transparency/. [305]

GIFCT (n.d.), *About our Leadership*, https://gifct.org/leadership/. [313]

GIFCT (n.d.), *Global Internet Forum to Counter Terrorism: Evolving an Institution*, https://gifct.org/about/. [13]

GIFCT (n.d.), *Join Tech Innovation*, https://gifct.org/joint-tech-innovation/. [300]

GIFCT Staff (2021), *Conclusion & Initial Next Steps*. [307]

GIFCT Transparency Working Group (2021), *One-year Review of Discussions*. [46]

Google (2019), *Google Drive Terms of Service*, https://www.google.com/drive/terms-of-service/. [208]

Google (n.d.), *Abuse program policies and enforcement - Docs Editors Help*, https://support.google.com/docs/answer/148505?visit_id=637064013896463652-1393240150&rd=1. [206]

Google (n.d.), *Google Transparency Report*, https://transparencyreport.google.com/?hl=en_GB. [211]

Google (n.d.), *Google Transparency Report*, https://transparencyreport.google.com/?hl=en_GB. [124]

Google (2010-2021), *Government requests to remove content - Google Transparency Report*, https://transparencyreport.google.com/government-removals/overview?hl=en_GB. [125]

Google (n.d.), *Report a violation - Docs Editors Help*, https://support.google.com/docs/answer/2463296?hl=en&ref_topic=1360897. [210]

Google (n.d.), *Request a review of a violation - Docs Editors Help*, https://support.google.com/docs/answer/2463328?hl=en&ref_topic=1360897. [207]

Google, Youtube (2017-2020), *Google Transparency Report - Flags*, https://transparencyreport.google.com/youtube-policy/flags?request_examples=year:;flagging_reason:7;flagger_type:&lu=request_examples. [304]

Google/YouTube (2021), *Appeal Community Guidelines actions*, https://support.google.com/youtube/answer/185111?hl=en. [114]

Google/YouTube (2021), *Community Guidelines - How does YouTube identify content that violates the Community Guidelines?*, https://www.youtube.com/intl/ALL_in/howyoutubeworks/policies/community-guidelines/#detecting-violations. [116]

Google/YouTube (2021), *Community Guidelines - What action does YouTube take for content that violates the Community Guidelines?*, https://www.youtube.com/intl/ALL_in/howyoutubeworks/policies/community-guidelines/#taking-action-on-violations. [118]

Google/YouTube (2021), *Community Guidelines strike basics - YouTube Help*, https://support.google.com/youtube/answer/2802032. [122]

Google/YouTube (2021), *Disable or enable Restricted/Safe Mode*, https://support.google.com/youtube/answer/174084?hl=en. [115]

Google/YouTube (2021), *Limited features for certain videos - YouTube Help*, https://support.google.com/youtube/answer/7458465. [123]

Google/YouTube (2021), *YouTube Community Guidelines enforcement - Violent Extremism*, https://transparencyreport.google.com/youtube-policy/featured-policies/violent-extremism?hl=en_GB&policy_removals=period:Y2019Q2&lu=policy_removals. [119]

Google/YouTube (2021), *YouTube Community Guidelines Enforcement FAQs*, https://support.google.com/transparencyreport/answer/9209072#zippy=%2Chow-is-violative-view-rate-vvr-calculated. [126]

Google/YouTube (2020), *YouTube Trusted Flagger program*, https://support.google.com/youtube/answer/7554338?&ref_topic=2803138. [117]

Google/Youtube (2019), *Our ongoing work to tackle hate*, https://blog.youtube/news-and-events/our-ongoing-work-to-tackle-hate. [113]

Gorwa, R. (2019), "The platform governance triangle: conceptualising the informal regulation of online content", *Internet Policy Rview*, Vol. 8/2. [65]

Government of Canada (2021), *Regulation of social media platforms*, https://search.open.canada.ca/en/qp/id/pch,PCH-2020-QP-00084?wbdisable=true. [292]

Government of Canada (2019), *Canada's Digital Charter: Trust in a Digital World*, https://www.ic.gc.ca/eic/site/062.nsf/eng/h_00108.html. [69]

Grüll, P. (2020), *German online hate speech reform criticised for allowing 'backdoor' data collection*, https://www.euractiv.com/section/data-protection/news/german-online-hate-speech-reform-criticised-for-allowing-backdoor-data-collection/. [263]

Guillemin, G. (2020), *EU Terrorist Content Regulation Rights Sell-Out*, https://medium.com/@gabrielleguillemin/eu-terrorist-content-regulation-rights-sell-out-f982561d670d. [19]

Harwell, D. and T. Romm (2019), *TikTok's Beijing roots fuel censorship suspicion as it builds a huge U.S. audience*, https://www.washingtonpost.com/technology/2019/09/15/tiktoks-beijing-roots-fuel-censorship-suspicion-it-builds-huge-us-audience/. [262]

Hatmaker, T. (2019), *This led to Reddit administrators banning the entire community in question from the site.*, https://techcrunch.com/2019/03/15/reddit-watchpeopledie-subreddit-gore/. [160]

Hayden, M. (2019), *Far-Right Extremists Are Calling for Terrorism on the Messaging App Telegram*, https://www.splcenter.org/hatewatch/2019/06/27/far-right-extremists-are-calling-terrorism-messaging-app-telegram. [152]

Hayden, M. (2019), *Mysterious Neo-Nazi Advocated Terrorism for Six Years Before Disappearance*, https://www.splcenter.org/hatewatch/2019/05/21/mysterious-neo-nazi-advocated-terrorism-six-years-disappearance. [205]

Hern, A. (2019), *Revealed: how TikTok censors videos that do not please Beijing*, https://www.theguardian.com/technology/2019/sep/25/revealed-how-tiktok-censors-videos-that-do-not-please-beijing. [261]

HM Government (2019), *Online Harms White Paper*, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf (accessed on 4 June 2019). [72]

HM Government (April 2019), *Online Harms White Paper*. [5]

Huang, F. (2018), *China's Most Popular App Is Full of Hate*, https://foreignpolicy.com/2018/11/27/chinas-most-popular-app-is-full-of-hate/. [143]

Huqing, W. and S. Zhixin (2013), "Research on Zero-Knowledge Proof Protocol", *International Journal of Computer Science Issues*, Vol. 10/1. [317]

Hymas, C. (2019), *Isil extremists using Instagram to promote jihad and incite support for terror attacks on the West*, https://www.telegraph.co.uk/news/2019/05/11/isil-extremists-using-instagram-promote-jihad-incite-support/. [136]

Ilinsky, A. (2019), *Interview with Mariusz Zurawek, founder of JustPaste.it, the anonymous sharing tool*, https://hostadvice.com/blog/justpaste-it-is-the-quickest-way-to-share-content-online/. [64]

IMO (n.d.), *IMO Community Guidelines*, https://imo.im/policies/community_guidelines.html. [176]

ISDGlobal (n.d.), *Powering solutions to extremism and polarisation*, https://www.isdglobal.org/. [121]

Jasser, G. and J. McSwiney (2021), "'Welcome to #GabFam': Far- right virtual community on Gab", *New Media & Society*, Vol. 18/1. [25]

JOYY Inc. (2021), *JOYY Reports First Quarter 2021 Unaudited Financial Results*, https://ir.joyy.sg/node/9156/pdf. [93]

Justepaste.it (2018), *Anonymous by default*, https://justpaste.it/1ikbg. [63]

Justpaste.it (2021), *JustPaste.it Transparency Report 2020*, https://justpaste.it/transparency_report_2020. [231]

Justpaste.it (2020), *JustPaste.it Transparency Report 2019*, https://justpaste.it/transparency_report_2019. [232]

Justpaste.it (n.d.), *JustPaste.it - Share Text & Images the Easy Way*, https://justpaste.it/about. [315]

Kakao (n.d.), *Operation Policy*, https://www.kakao.com/policy/oppolicy?lang=en. [198]

Karadeglija, A. (2021), *New definition of hate to be included in Liberal bill that might also revive contentious hate speech law*, https://nationalpost.com/news/politics/new-definition-of-hate-to-be-included-in-liberal-bill-that-might-also-revive-contentious-hate-speech-law. [260]

Kastrenakes, J. (2021), *Apple says there are now over 1 billion active iPhones*, https://www.theverge.com/2021/1/27/22253162/iphone-users-total-number-billion-apple-tim-cook-q1-2021. [75]

Katz, R. (2020), *Neo-Nazis Are Running Out of Places to Hide Online*, https://www.wired.com/story/neo-nazis-are-running-out-of-places-to-hide-online/. [31]

Katz, R. (2019), *A Growing Frontier for Terrorist Groups: Unsuspecting Chat Apps*, https://www.wired.com/story/terrorist-groups-prey-on-unsuspecting-chat-apps/. [145]

Katz, R. (2019), *Opinion: Telegram has finally cracked down on Islamist terrorism. Will it do the same for the far-right?*, https://www.washingtonpost.com/opinions/2019/12/05/telegram-has-finally-cracked-down-islamist-terrorism-will-it-do-same-far-right/. [30]

Katz, R. (2018), *To Curb Terrorist Propaganda Online, Look to YouTube. No, Really.*, https://www.wired.com/story/to-curb-terrorist-propaganda-online-look-to-youtube-no-really/. [212]

Keller, D. and P. Leerssen (2020), *Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation*, Cambridge University Press. [17]

Kemp, S. (2021), *Digital 2021: Global Overview Report*, https://datareportal.com/reports/digital-2021-global-overview-report. [258]

Kemp, S. (2019), *Digital 2019: Global Digitial Overview*, https://datareportal.com/reports/digital-2019-global-digital-overview. [259]

Kenny, K. (2019), *How can upcoming social media efforts be 'global' if they ignore Asia?*, https://www.stuff.co.nz/national/christchurch-shooting/112284082/how-can-upcoming-social-media-efforts-be-global-if-they-ignore-asia. [142]

Kenyon, M. (2020), *WeChat Surveillance Explained*, https://citizenlab.ca/2020/05/wechat-surveillance-explained/. [257]

King, J. (2020), *How We Review Content*, https://about.fb.com/news/2020/08/how-we-review-content/. [106]

Kitsune, L. (2017), *New Notifications and Reporting Updates by Lauren Kitsune on DeviantArt*, https://www.deviantart.com/laurenkitsune/journal/New-Notifications-and-Reporting-Updates-706864447. [200]

Knockel, J. et al. (2020), *We Chat, They Watch How International Users Unwittingly Build up WeChat's Chinese Censorship Apparatus*, https://citizenlab.ca/2020/05/we-chat-they-watch/. [141]

Knockel, J. et al. (2018), *(Can't) Picture This, An Analysis of Image Filtering on WeChat Moments*, https://citizenlab.ca/2018/08/cant-picture-this-an-analysis-of-image-filtering-on-wechat-moments/. [53]

Kock, R. (2020), *TikTok and the privacy perils of China's first international social media platform*, 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. [256]

Kulvi, F. (2021), *Meet Rumble, Canada's new 'free speech' platform — and its impact on the fight against online misinformation*, https://theconversation.com/meet-rumble-canadas-new-free-speech-platform-and-its-impact-on-the-fight-against-online-misinformation-163343. [235]

Lange, D. (2017), *Quora's Tolerance Of Terror Support*, https://www.israellycool.com/2017/05/22/quoras-tolerance-of-terror-support/. [171]

Lardinois, F. (2020), *Microsoft Teams is coming to consumers — but Skype is here to stay*, https://techcrunch.com/2020/03/30/microsoft-teams-is-coming-to-consumers-but-skype-is-here-to-stay/. [94]

Le Monde (2021), *Haine en ligne : des obligations de transparence pour les réseaux sociaux*, https://www.lemonde.fr/politique/article/2021/01/18/haine-en-ligne-des-obligations-de-transparence-pour-les-reseaux-sociaux_6066656_823448.html. [302]

Li Xan Wong, K. and A. and Shields Dobson (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*, Vol. 4/2, pp. 220-232. [255]

Liao, R. (2019), *PicsArt hits 130 million MAUs as Chinese flock to its photo-editing app*, https://techcrunch.com/2019/03/20/picsart-china/. [184]

Liao, S. (2018), *Discord shuts down more neo-Nazi, alt-right servers*, https://www.theverge.com/2018/2/28/17062554/discord-alt-right-neo-nazi-white-supremacy-atomwaffen. [185]

LINE (2020), *LINE Content Moderation Report*, https://linecorp.com/en/security/moderation/2019h1. [180]

LINE (n.d.), *Help Center*, https://help.line.me/line/android/categoryId/20000132/3/pc?lang=en. [181]

Line Corporation (2020), *LINE Q3 2020 Earnings Results*, https://d.line-scdn.net/stf/linecorp/en/ir/all/FY20Q3_earning_releases_EN.pdf. [89]

Lix Xan Wong, K. and A. Shields Dobson (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*, Vol. 4/2, pp. 220-232. [57]

Lu, S. (2021), *I helped build ByteDance's vast censorship machine*, https://www.protocol.com/china/i-built-bytedance-censorship-machine. [59]

Lymn, T. (2021), *The use of algorithms in the content moderation process*, https://cdei.blog.gov.uk/2021/08/05/the-use-of-algorithms-in-the-content-moderation-process/. [16]

Mail.ru (2021), *Social Networks*, https://corp.mail.ru/en/company/social/. [95]

Malik, N. (2018), *Terror in the Dark: How Terrorists Use Encryption, the Darknet, and Cryptocurrencies*, The Henry Jackson Society. [254]

Marketing to China (2021), *Guide to Douban Marketing*, https://marketingtochina.com/guide-to-douban-marketing/. [82]

Marketing to China (2021), *Top 10 Chinese Social Media for Marketing (updated 2021)*, https://www.marketingtochina.com/top-10-social-media-in-china-for-marketing/. [83]

May, T. (2018), *Theresa May's Davos address in full*, https://www.weforum.org/agenda/2018/01/theresa-may-davos-address/. [7]

Medium (2015), *Medium's Transparency Report (2014)*, https://medium.com/transparency-report/mediums-transparency-report-438fe06936ff. [179]

Medium (n.d.), *Controversial, Suspect and Extreme Content, Medium Help Center*, https://help.medium.com/hc/en-us/articles/360018182453. [178]

Mega.nz (2021), *Mega Transparency Report - September 2021*, https://mega.io/Mega_Transparency_Report_September_2021.pdf. [62]

Mega.nz (2021), *Mega Transparency Report 2021 (blogpost)*, https://mega.io/blog/mega-transparency-report-2021. [61]

Meleagrou-Hitchens, A. and N. Kaderbhai (2017), *Research Perspectives on Online Radicalisation - A Literature Review, 2006-2016*. [253]

Microsoft (2021), *Digital Safety Content Report*, https://www.microsoft.com/en-us/corporate-responsibility/digital-safety-content-report?activetab=pivot_1:primaryr4. [306]

Microsoft (2016), *Microsoft's approach to terrorist content online, Microsoft on the Issues*, https://blogs.microsoft.com/on-the-issues/2016/05/20/microsofts-approach-terrorist-content-online/#sm.000del1ea19zbe4duja1ve96fcc1l. [172]

Microsoft (2020-2021), *Digital Safety Content Report*, https://www.microsoft.com/en-us/corporate-responsibility/digital-safety-content-report?activetab=pivot_1:primaryr4. [175]

Microsoft (n.d.), *Report abuse in Teams*, https://support.microsoft.com/en-us/office/report-abuse-in-teams-2e2ea20c-2866-4b65-a979-8132c02dc231. [174]

Miller, J. (2014), *Can Iraqi militants be kept off social media sites?*. [177]

Odnoklassniki (n.d.), *Help Centre*, https://ok.ru/help/54/367. [193]

Odysee (2021), *Moderation Tools*, https://odysee.com/@OdyseeHelp:b/moderation:f. [237]

OECD (2021), *Transparency Reporting - Considerations for the Review of the Privacy Guidelines*. [45]

OECD (2021), *Transparency Reporting on Terrorist and Violent Extremist Content Online - An Update on the Global Top 50 Content-sharing services*, OECD Publishing, Paris. [15]

OECD (2020), *Current Approaches to Terrorist and Violent Extremist Content Among the Global Top 50 Online Content-sharing Services*, OECD Publishing, Paris. [14]

OECD (2017), *Economic and Social Benefits of Internet Openness - 2016 Ministerial Meeting on the Digital Economy, Background Report*. [1]

Parler (2021), *Community Guidelines*, https://parler.com/documents/guidelines.pdf. [236]

Parliament of the Commonwealth of Australia, House of Representatives (2021), *Explanatory Memorandum to the Online Safety Act*, https://parlinfo.aph.gov.au/parlInfo/download/legislation/ems/r6680_ems_3499aa77-c5e0-451e-9b1f-01339b8ad871/upload_pdf/JC001336%20Clean4.pdf;fileType=application%2Fpdf. [66]

Patriquin, M. (2021), *With new legislation, Steven Guilbeault will make few friends in Big Tech*, https://financialpost.com/technology/with-new-legislation-steven-guilbeault-will-make-few-friends-in-big-tech. [252]

Penetrum Security (n.d.), *Petrum Security Analysis of TikTok versions 10.0.8 -15.2.3*. [296]

Perez, S. (2020), *TikTok to open a 'Transparency Center' where outside experts can examine its content moderation practices*, https://techcrunch.com/2020/03/11/tiktok-to-open-a-transparency-center-where-outside-experts-can-examine-its-moderation-practices/. [250]

Perez, S. (2019), *Skype publicly launches screen sharing on iOS and Android*, https://techcrunch.com/2019/06/05/skype-publicly-launches-screen-sharing-on-ios-and-android/?guccounter=1. [251]

Picsart (2015), *Picsart Community Guidelines Blogpost*, https://picsart.com/blog/post/picsart-community-guidlines. [183]

Picsart (n.d.), *How do I report inappropriate behavior or content on Picsart?*, https://support.picsart.com/hc/en-us/articles/360003824257-How-do-I-report-inappropriate-behavior-or-content-on-Picsart-. [182]

Pinterest (n.d.), *Account suspension*, https://help.pinterest.com/en/article/account-suspension. [155]

Powell, B. (2019), *Encrypted Extremism - Inside the English-Speaking Islamic State Ecosystem on Telegram*. [195]

Prucha, N. (2016), "IS and the Jihadist Information Highway – Projecting Influence and Religious Identity via Telegram", *Perspectives on Terrorism*, Vol. 10/6. [23]

Quay-de la Vallee, H. and M. Azarmi (2020), *The New EARN IT Act Still Threatens Encryption and Child Exploitation Prosecutions*, https://cdt.org/insights/the-new-earn-it-act-still-threatens-encryption-and-child-exploitation-prosecutions/. [249]

Quinn, B. (2021), *Telegram is warned app 'nurtures subculture deifying terrorists'*, https://www.theguardian.com/uk-news/2021/oct/14/telegram-warned-of-nurturing-subculture-deifying-terrorists. [32]

Quora (n.d.), *How does Quora Moderation make decisions about edit-blocks and bans?*, https://help.quora.com/hc/en-us/articles/360001069906-How-does-Quora-Moderation-make-decisions-about-edit-blocks-and-bans-. [170]

Ranking Digital Rights - Tencent Holdings Limited (2021), *2020 Ranking Digital Rights Corporate Accountability Index*, https://rankingdigitalrights.org/index2020/companies/Tencent#ftnt2. [139]

Rasmussen, N. and J. Lowin (2021), *Introduction*. [37]

Reddit (2022), *Transparency Report 2021*, https://www.redditinc.com/policies/transparency-report-2021. [156]

Reddit (2020), , https://www.redditinc.com/policies/transparency-report-2020-1. [276]

Reddit (2020), *Transparency Report 2020*, https://www.redditinc.com/policies/transparency-report-2020-1. [279]

Reddit Inc. (2020), *Transparency Report 2020*, https://www.redditinc.com/policies/transparency-report-2020-1. [278]

Reddit Inc. (2017), *Moderator Guidelines for Healthy Communities*, https://www.redditinc.com/policies/moderator-guidelines. [158]

Reddit Inc. (n.d.), *AutoModerator*, https://mods.reddithelp.com/hc/en-us/articles/360002561632-AutoModerator. [159]

Reddit Inc. (n.d.), *Quarantined Subreddits*, https://www.reddithelp.com/en/categories/rules-reporting/account-and-community-restrictions/quarantined-subreddits. [157]

Rosenthal, M. (ed.) (2014), *Vkontakte, a Russian social network, is hosting ISIS accounts that were kicked off of Facebook and Twitter*, https://www.pri.org/stories/2014-09-12/isis-internet-army-has-found-safe-haven-russian-social-networks-now. [192]

Ruan, L. (2019), *Regulation of the internet in China: An explainer*, https://theasiadialogue.com/2019/10/07/regulation-of-the-internet-in-china-an-explainer/. [148]

Ruan, L. (2016), *One App, Two Systems, How WeChat uses one censorship policy in China and another internationally*, https://citizenlab.ca/2016/11/wechat-china-censorship-one-app-two-systems/. [58]

Santa Clara University's High Tech Law Institute (n.d.), *The Santa Clara Principles On Transparency and Accountability in Content Moderation*, https://santaclaraprinciples.org/. [52]

Scott, M. and L. Kayali (2020), *What happened when humans stopped managing social media content*, https://www.politico.eu/article/facebook-content-moderation-automation/. [50]

Scrivens, R. et al. (2021), "Examining Online Indicators of Extremism in Violent Right-Wing Extremist Forums", *Studies in Conflict & Terrorism*. [27]

Sensor Tower (2020), *How PicsArt has Thrived in the Competitive Photo & Video Category*, https://sensortower.com/blog/picsart-interview. [90]

Silver, J. (2021), *Regulation of online hate speech coming soon, says minister*, https://ipolitics.ca/2021/01/29/regulation-of-online-hate-speech-coming-soon-says-minister/. [248]

Similarweb (2021), *vk.com*, https://www.similarweb.com/website/vk.com/. [273]

Site Intelligence Group Enterprise (2018), *IS-linked Media Group Makes Foray onto Viber Messenger - Dark Web and Cyber Security*, https://ent.siteintelgroup.com/Dark-Web-and-Cyber-Security/is-linked-media-group-makes-foray-onto-viber-messenger.html. [144]

Sivakumar, B. (2020), *YouTube vs Vimeo: A Detailed Comparison*, https://www.feedough.com/youtube-vs-vimeo/. [247]

Sky News (2020), *FBI unlocks terrorist's iPhones and finds al Qaeda links - 'no thanks to Apple'*, https://news.sky.com/story/fbi-unlocks-terrorists-iphones-and-finds-al-qaeda-links-no-thanks-to-apple-11990818. [134]

Smith, C. (2021), *IMO Statistics, User Counts and Facts (2021)*, https://expandedramblings.com/index.php/imo-facts-and-statistics/. [85]

Snap Inc. (2021), *Transparency Report*, https://snap.com/en-GB/privacy/transparency. [154]

Snap Inc. (n.d.), *Safety Centre, Report a safety concern*, https://www.snap.com/en-GB/safety/safety-reporting/. [153]

START (National Consortium for the Study of Terrorism and Responses to Terrorism) (2018), *The Use of Social Media by United States Extremists*, https://www.start.umd.edu/pubs/START_PIRUS_UseOfSocialMediaByUSExtremists_ResearchBrief_July2018.pdf. [169]

Startup Talky (2020), *YouTube vs Vimeo: A Detailed Comparison*, https://startuptalky.com/youtube-vs-vimeo/. [87]

Statista (2021), *Average number of monthly active users (MAU) of Chinese video app iQiyi from 2016 to 2020*, https://www.statista.com/statistics/1106091/china-online-video-platform-iqiyi-mobile-app-monthly-active-user-number/. [79]

Statista (2021), *Number of global monthly active Kakaotalk users from 1st quarter 2013 to 4th quarter 2020*, https://www.statista.com/statistics/278846/kakaotalk-monthly-active-users-mau/. [98]

Statista (2021), *Number of monthly active users of popular short video apps in China as of May 2021*, https://www.statista.com/statistics/910633/china-monthly-active-users-across-leading-short-video-apps/. [96]

Statista (2021), *Total global visitor traffic to Zoom.us 2021*, https://www.statista.com/statistics/1259905/zoom-website-traffic/. [74]

Stokel-Walker, C. (2020), *As humans go home, Facebook and YouTube face a coronavirus crisis*, https://www.wired.co.uk/article/coronavirus-facts-moderators-facebook-youtube. [246]

Tardi, C. (2020), *Monthly Active Users (MAU)*, https://www.investopedia.com/terms/m/monthly-active-user-mau.asp. [316]

Tech Against Terrorism (2021), *GIFCT Technical Approaches Working Group - Gap Analysis and Recommendations for deploying technical solutions to tackle the terrorist use of the Internet*. [29]

Tech Against Terrorism (2020), *The Online Regulation Series - India*, https://www.techagainstterrorism.org/2020/10/09/the-online-regulation-series-india/. [286]

Tech Against Terrorism (2020), *The Online Regulation Series - Singapore*, https://www.techagainstterrorism.org/2020/10/05/the-online-regulation-series-singapore/. [285]

Tech Against Terrorism (2020), *The Online Regulation Series: The United States*, https://www.techagainstterrorism.org/2020/10/13/the-online-regulation-series-the-united-states/. [284]

Tech Against Terrorism (2019), *Analysis: ISIS use of smaller platforms and the DWeb to share terrorist content – April 2019*, https://www.techagainstterrorism.org/2019/04/29/analysis-isis-use-of-smaller-platforms-and-the-dweb-to-share-terrorist-content-april-2019/. [22]

Tech Against Terrorism (2019), *Analysis: ISIS use of smaller platforms and the DWeb to share terrorist content – April 2019*, https://www.techagainstterrorism.org/2019/04/29/analysis-isis-use-of-smaller-platforms-and-the-dweb-to-share-terrorist-content-april-2019/. [311]

Techcircle (2021), *Microsoft Teams reaches 250 million global MAU milestone*, https://www.techcircle.in/2021/07/29/microsoft-teams-reaches-250-million-global-mau-milestone. [84]

Telegram (n.d.), *ISIS Watch*, https://telegram.me/ISISwatch. [150]

Telegram (n.d.), *Telegram Privacy Policy*, https://telegram.org/privacy. [149]

The American University in Cairo (n.d.), *Terrorism vs. Extremism: Are They Linked?*, https://www.aucegypt.edu/news/stories/terrorism-vs-extremism-are-they-linked. [288]

The Hindu Business Line (2020), *Social media users to be tracked by government under new guidelines*, https://www.thehindubusinessline.com/info-tech/social-media/social-media-users-to-be-tracked-by-government-under-new-guidelines-report/article30807839.ece. [289]

The International Centre for the Study of Radicalisation (ICSR) (2020), *ICSR info*, https://icsr.info/. [120]

The Santa Clara Principles (n.d.), *The Santa Clara Principles on Transparency and Accountability in Content Moderation*, https://santaclaraprinciples.org. [283]

Thomsom, E. (2021), *Canada not exempt from social media forces that created U.S. Capitol riot, heritage minister says*, https://www.cbc.ca/news/politics/facebook-twitter-canada-regulation-1.5894301. [245]

Thune, J. (2020), *Thune: PACT Act Would Increase Internet Accountability and Consumer Transparency*, https://www.thune.senate.gov/public/index.cfm/2020/7/thune-pact-act-would-increase-internet-accountability-and-consumer-transparency. [244]

TikTok (2022), *TikTok Transparency Report*, https://www.tiktok.com/transparency/en-us/community-guidelines-enforcement-2021-3/. [146]

TikTok (2021), *Community Guidelines*, https://www.tiktok.com/community-guidelines?lang=en#39. [41]

TikTok (2019-2020), *TikTok Transparency Report*, https://www.tiktok.com/safety/resources/transparency-report?lang=en. [147]

Titcomb, J. (2017), *Why Google is reading your Docs*, https://www.telegraph.co.uk/technology/2017/11/01/google-reading-docs/. [209]

Trentmann, N. and S. Needleman (2021), *Chat Startup Discord Hires Its First Finance Chief to Boost Growth*, https://www.wsj.com/articles/chat-startup-discord-hires-its-first-finance-chief-to-boost-growth-11616086771. [243]

Trudeau, J. (2021), *Minister of Canadian Heritage Mandate Letter*, https://pm.gc.ca/en/mandate-letters/2021/12/16/minister-canadian-heritage-mandate-letter. [70]

Trudeau, J. (2021), *Minister of Justice and Attorney General of Canada Mandate Letter*, https://pm.gc.ca/en/mandate-letters/2021/12/16/minister-justice-and-attorney-general-canada-mandate-letter. [71]

Tumblr (2013-2020), *Tumblr Transparency Report*, https://www.tumblr.com/transparency. [166]

Twitch (2021), *Hateful Conduct and Harassment*, https://www.twitch.tv/p/en/legal/community-guidelines/harassment/. [39]

Twitch (2021), *Transparency Report*, https://www.twitch.tv/p/en/legal/transparency-report/. [49]

Twitch (2020), *Transparency Report 2020*, https://www.twitch.tv/p/en/legal/transparency-report/. [277]

Twitter (2021), *Rules Enforcement*, https://transparency.twitter.com/en/reports/rules-enforcement.html#2020-jul-dec. [290]

Twitter (2020), *Updating our rules against hateful conduct*, https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate. [40]

Twitter (n.d.), *Our approach to policy development and enforcement philosophy*, https://help.twitter.com/en/rules-and-policies/enforcement-philosophy. [162]

Twitter (n.d.), *Our range of enforcement options*, https://help.twitter.com/en/rules-and-policies/enforcement-options. [161]

Twitter (2012-2021), *Twittter Rules enforcement*, https://transparency.twitter.com/en/twitter-rules-enforcement.html. [163]

United Nations Office on Drugs and Crime (2012), *The use of the Internet for terrorist purposes*, United Nations Office at Vienna. [282]

United Nations Security Council (n.d.), *United Nations Security Council Consolidated List*, https://www.un.org/securitycouncil/content/un-sc-consolidated-list. [173]

US Treasury (2020), *OFFICE OF FOREIGN ASSETS CONTROL - Specially Designated Nationals and Blocked Persons List*, https://www.treasury.gov/ofac/downloads/sdnlist.pdf. [217]

V-click Technology (2021), *Stats summary about China social media in 2020 when covid-19 this the world*, https://www.v-click.co.th/china-social-media-in-2020-covid/. [77]

Verizon Media (2019), *Transparency Report*, https://www.verizonmedia.com/transparency/index.html?guce_referrer=aHR0cHM6Ly90cmFuc3BhcmVuY3kub2F0aC5jb20vaW5kZXguaHRtbD9ndWNlX3JlZmVycmVyPWFIUjBjSE02THk5M2QzY3VkSFZ0WW14eUxtTnZiUzgmZ3VjZV9yZWZlcnJlcl9zaWc9QVFBQUFKazduZ3VNWS04dHhtNG9hWFM3TUlkNkxlUWxkMEZ5. [165]

VK (2021), *Healthy Environment*, https://m.vk.com/safety?section=health. [191]

VK (2021), *Platform Standards*, https://m.vk.com/safety?section=standards. [42]

VK (2020), *Platform Standards*, https://m.vk.com/safety?lang=en&section=standarts. [295]

VK (2020), *Social Accountability*, https://m.vk.com/safety?section=social&lang=en. [190]

Volkmer, B. (2019), *Protecting our users and society: guarding against terrorist content*, https://blog.dropbox.com/topics/company/protecting-our-users-and-society--guarding-against-terrorist-con. [213]

Wahlström, M., A. Törnberg and E. Hans (2021), "Dynamics of violent and dehumanizing rhetoric in far-right social media", *New media & society*, Vol. 23/11. [314]

Wang, M. (2019), *Wechatscope*, https://advox.globalvoices.org/2019/02/11/censored-on-wechat-a-year-of-content-removals-on-chinas-most-powerful-social-media-platform/. [56]

Warner, A. (2021), *Which Social Media Platform Has the Most Users?*, https://www.websiteplanet.com/blog/social-media-platform-users/. [78]

WeGo.Social (n.d.), *About WeGo.Social*, https://wego.social/terms/about-us. [238]

Weimann, G. (2016), *Why do terrorists migrate to social media?*, Abingdon: Routledge. [4]

Weimann, G. (2014), *New Terrorism and New Media*, Commons Lab of the Woodrow Wilson International Center for Scholars, https://www.wilsoncenter.org/sites/default/files/new_terrorism_v3_1.pdf. [197]

WhatsApp (2021), *Terms of Service*, https://www.whatsapp.com/legal/terms-of-service/?lang=en. [129]

Wikimedia (2021), *Unique devices*, https://www.envisagedigital.co.uk/wordpress-market-share/. [103]

Wikimedia Foundation (2019), *Terms of Use - Wikimedia Foundation Governance Wiki*, https://foundation.wikimedia.org/wiki/Terms_of_Use/en. [229]

Wikimedia Foundation (n.d.), *Transparency report*, https://transparency.wikimedia.org/. [230]

Wikipedia (2020), *Administration - Wikipedia*, https://en.wikipedia.org/wiki/Wikipedia:Administration#Human_and_legal_administration. [224]

Wikipedia (2020), *Criteria for speedy deletion - Wikipedia*, https://en.wikipedia.org/wiki/Wikipedia:Criteria_for_speedy_deletion#Procedure_for_administrators. [228]

Wikipedia (2020), *Deletion process - Wikipedia*, https://en.wikipedia.org/wiki/Wikipedia:Deletion_process. [227]

Wikipedia (2020), *Oversight - Wikipedia*, https://en.wikipedia.org/wiki/Wikipedia:Oversight. [223]

Wikipedia (2020), *What Wikipedia is not - Wikipedia*. [226]

Wikipedia (2019), *CheckUser - Wikipedia*, https://en.wikipedia.org/wiki/Wikipedia:CheckUser. [222]

Wikipedia (2019), *Core Content Policies - Wikipedia*, https://en.wikipedia.org/wiki/Wikipedia:Core_content_policies. [225]

Willens, M. (2021), *Four years into a subscription strategy, Medium still doesn't spend money to acquire subscribers*, https://digiday.com/media/four-years-into-a-subscription-strategy-medium-still-doesnt-spend-money-to-acquire-subscribers/. [88]

Wimkin (2021), *Wimkin Takes User Safety Extremely Seriously*, https://wimkinhelp.com/report-threats. [240]

Word Press (n.d.), *Terrorist Activity - Support - Word Press.com*, https://en.support.wordpress.com/terrorist-activity/. [218]

WordPress.com (n.d.), *Suspended Content and Sites*, https://en.support.wordpress.com/suspended-blogs/. [220]

Worldtruthvideos.website (2022), *About us*, https://worldtruthvideos.website/terms/about-us. [241]

Yahoo! Finance (2019), *YY earnings surpass estimates in Q3, revenues increase*, https://finance.yahoo.com/news/yy-earnings-surpass-estimates-q3-144502223.html. [272]

YouTube (2020), *Protecting our extended workforce and the community*, https://blog.youtube/news-and-events/protecting-our-extended-workforce-and. [293]

Zetter, K. (2015), *Security Manual Reveals the OPSEC Advice ISIS Gives Recruits*, https://www.wired.com/2015/11/isis-opsec-encryption-manuals-reveal-terrorist-group-security-protocols/. [132]

Zhang, Y. (2018), *Global Times*, http://www.globaltimes.cn/content/1098173.shtml. [55]

Zhong, R. (2018), *At China's Internet Conference, a Darker Side of Tech Emerges*, https://www.nytimes.com/2018/11/08/technology/china-world-internet-conference.html. [140]

Zoom (2021), *Community Standards Enforcement*, https://explore.zoom.us/en/trust/community-standards-enforcement/. [127]

Zoom (2021), *Out Tier Review System*, https://explore.zoom.us/docs/en-us/content-moderation-process.html?_ga=2.20044602.38595736.1624527871-1107759908.1602261224. [128]

Zoom (2020-2021), *Transparency Report*, https://explore.zoom.us/docs/en-us/trust/transparency-12-18-2020.html?_ga=2.83383058.50725218.1634056115-184629636.1634056115. [281]

Endnotes

[1] Available at https://www.oecd-ilibrary.org/docserver/5jlwqf2r97g5-en.pdf?expires=1635859372&id=id&accname=guest&checksum=7226C3F9479EFE77323EAD4D6FC34229

[2] "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, 2020[316]).

[3] Archived and contextualized means stored in an organised manner and complemented with metadata and other information indicating that the content is, and was intended to be, TVEC (as opposed to, e.g., journalism or educational material).

[4] See Section 1 of the Parler profile in Annex D.

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

[6] 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.

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

[8] Facebook, YouTube, TikTok, Twitch, Twitter and Google Drive profiles.

[9] See Section 1 of the profiles of Facebook, YouTube, Zoom, Instagram, Facebook Messenger, TikTok, Twitter, Vimeo, Picsart, Discord and Likee. Arguably, Microsoft (LinkedIn, Teams, Skype and OneDrive) belongs in this group, as well, though it provides no definition of violent extremism and does not offer any examples. Similarly, Google Drive has an explicit prohibition of terrorist content, but the definition revolves around conduct by terrorist organisations, which are not defined. Also, Pinterest provides good descriptions of hateful activities and content, but it does not define extremists and terrorist organisations.

[10] 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.

[11] 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.

[12] See Section 1 of the profiles of WeChat, Snapchat, Pinterest, Tumblr, LinkedIn, Quora, Teams, IMO, Ask.fm, Twitch, Skype, VK, Xigua Video, Flickr, Huoshan, Google Drive, Dropbox, OneDrive and Wordpress.com.

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

[14] WeChat, Instagram, QQ, Youku Tudou, iQIYI, Kuaishou, Douban, LinkedIn, Baidu Tieba, Vimeo, Medium, Odnoklassniki, and Meetup profiles.

[15] See Section 1 of the profiles of Kuaishou, iQIYI, Baidu Tieba, Medium and Odnoklassniki

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

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

[18] See Section 1 of the profiles of WhatsApp, iMessage/Facetime, Viber, QQ, Youku Tudou, Telegram, Qzone, Weibo, Reddit, Douban, LINE, Kakao, Smule, DeviantArt and Wikipedia.

[19] See Section 5 of the Tumblr profile.

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

[21] See Section 7 of the YouTube profile, and Section 1 of the Skype, Quora, Microsoft OneDrive and Wordpress.com profiles.

[22] 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, Kakao, DeviantArt, Meetup, 4chan, Google Drive, Dropbox and Wikipedia profiles.

[23] See Section 5 of Zoom's profile.

[24] Facebook, YouTube, WhatsApp, Facebook Messener, 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, Kakao, DeviantArt, Meetup, 4chan, MySpace, Google Drive, Dropbox, OneDrive, WordPress.com and Wikipedia.

[25] 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, Kakao, DeviantArt, Meetup, 4chan, Google Drive, Dropbox, OneDrive, WordPress.com and Wikipedia.

[26] See Section 5 of the profiles of Facebook, YouTube, Zoom, WhatsApp, iMessage/Facetime, Instagram, Facebook Messenger, WeChat, Viber, TikTok, QQ, Youku Tudou, Telegram, Qzone, Weibo, Snapchat, Kuaishou, iQIYI, Pinterest, Reddit, Twitter, Tumblr, LinkedIn, Douban, Baidu Tieba, Quora, Teams, IMO, Ask.fm, Vimeo, Medium, LINE, Picsart, Discord, Twitch, Likee, Skype, VK, Xigua Video, Odnoklassniki, Flickr, Huoshan, Kakao, Smule, DeviantArt, Google Drive, Dropbox, OneDrive, Wordpress.com, Wikipedia.

[27] Reddit, Viber, Twitch, Flickr, VK, Odnoklassniki, Kakao, DeviantArt, 4chan and Wikipedia.

[28] Reddit, Viber, Twitch, Flickr, VK, Odnoklassniki, Kakao, DeviantArt, 4chan and Wikipedia.

[29] See Section 5 of the profiles of Reddit, Viber, Discord, Twitch, VK, Odnoklassniki, Flickr, Kakao, DeviantArt and Wikipedia.

[30] 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.

[31] 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).

[32] 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.

[33] 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), Discord, 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).

[34] See Section 5 of the profiles of Facebook, YouTube, Zoom, WhatsApp, Instagram, Facebook Messenger, WeChat, Viber, TikTok, QQ, Youku Tudou, QZone, Weibo, Snapchat (Hash Sharing Consortium member), Kuaishou, iQIYI, Pinterest (Hash Sharing Consortium member), Reddit (Hash Sharing Consortium member), Twitter, Tumblr, LinkedIn, Douban, Baidu Tieba, Teams, IMO, Ask.fm, Vimeo, LINE, Picsart, Discord, Skype, Twitch, VK, Xigua Video, Odnoklassniki, Flickr, Huoshan, Kakao, DeviantArt, Google Drive, Dropbox, (GIFCT Member, ULR Sharing) and OneDrive.

[35] One of the GIFCT membership criteria is "the ability to receive, review, and act on both reports of activity that is illegal and/or violates terms of service", which presupposes having content moderators. Also, the database of the Hash Sharing Consortium is leveraged based on automated tools.

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

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

[38] See Section 4.1 of the profiles of Facebook, YouTube, Zoom, WhatsApp, Instagram, Facebook Messenger, WeChat, TikTok, Pinterest (at Pinterest's discretion), Reddit (account suspensions), Twitter, Tumblr (at Tumblr's discretion), LinkedIn, Quora (warnings), Teams, Ask.fm, Vimeo, Medium, Twitch, Skype, VK, Odnoklassniki, Flickr, Kakao, Smule (at Sumel's discretion), DeviantArt, Google Drive, Dropbox, OneDrive and Wordpress.com

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

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

[41] See Section 4.2 of the profiles of Facebook, YouTube, Zoom, WhatsApp, Instagram, Facebook Messenger, WeChat, Viber, TikTok, Kuaishou, Pinterest (account suspensions), Reddit, Twitter, Tumblr, LinkedIn, Quora (edit-blocks and

bans), Teams, Ask.fm, Vimeo, Medium, LINE, Discord, Twitch (warnings), Skype, VK, Kakao, DeviantArt, Google Drive, Dropbox, OneDrive and Wordpress.com

[42] 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.

[43] WhatsApp, iMessage/FaceTime, WeChat, Instagram, QQ, Weibo, iQIYI, Kuaishou, Douban, LinkedIn, Quora, Snapchat, Pinterest, IMO, Ask.fm, VK, Haokan, Odnoklassniki, Smule, Meetup, and OneDrive. Use of the word 'may' or the expression 'reserves the right to review', in particular, were very common.

[44] See Sections 4 and 5 of the profiles of iMessage/Facetime, Quora, Medium, Smule and Wikipedia.

[45] Figures do not add up to 25 in this column because one of the far-right-focused services – Thedonald.win – is no longer operational, so it was not possible to determine anything about its governing documents. Further information is provided in this Service's profile.

[46] See Section 1 of the YouTube, Twitter, Facebook, Instagram, TikTok and Discord profiles in Annex B.

[47] See Section 1 of the VK profile in Annex B.

[48] See Section 1 of the Google Drive and Dropbox profiles in Annex B and of Justpaste.it, Google Docs, Mega.nz and Pixeldrain in Annex D.

[49] See Section 1 of the BitChute, Rumble.com, Parler, Odysee and Doxbin profiles in Annex D.

[50] See Section 1 of the Telegram and WhatsApp profiles in Annex B and of the Element profile in Annex D.

[51] See Section 1 of the Archive.org, Files.fm, MediaFire, File.io, Gofile.io and Anonfiles profiles in Annex D.

[52] See Section 1 of the Gab, Brandnewtube, Gettr, 8kun, WeGoSocial, SafeChat, Wimkin, Worldtruthvideos and Xephula profiles in Annex D.

[53] See Section 1 of the Telegraph, Tlgur and Uploadgram profiles in Annex D.

[54] See Section 1 of the Patriots.win, Redvoicemedia.com, 88msn.com, Mzwnews.com, Thegreaterreset.org, Nordfront.dk, Lookaheadamerica.org, Patriotfront.us, Vastarinta.com and Nordicrresistancemovement.org profiles in Annex D.

[55] Figures do not add up to 25 in this column because one of the far-right-focused services – Thedonald.win – is no longer operational, so it was not possible to determine anything about its approach to content moderation. Further information is provided in this Service's profile

[56] See Section 5 of the Facebook, YouTube, Instagram, TikTok, Twitter, Discord and VK profiles in Annex B.

[57] See Section 5 of the Google Drive and Dropbox profiles in Annex B and of the Justpaste.it, MediaFire, Google Docs and Mega.nz profiles in Annex D.

[58] See Section 5 of the WhatsApp and Telegram profiles in Annex B and of the Element profile in Annex D.

[59] See Section 5 of the Files.fm and Pixeldrain profiles in Annex D.

[60] See Section 5 of the BitChute, Rumble.com, Parler, Odysee, 8kun, WeGoSocial and Wimkin profiles in Annex D.

[61] See Section 5 of the Gab, Gettr and SafeChat profiles in Annex D.

[62] See Section 5 of the Archive.org and Anonfiles profiles in Annex D.

[63] See Section 5 of the Doxbin and Xephula profiles in Annex D.

[64] See Section 5 of the Telegraph, Tlgur, Uploadgram, File.io and Gofile.io profiles in Annex D.

[65] See Section 5 of the Patriots.win, Brandnewtube, Redvoicemedia.com, 88nsm.com, Mzwnews.com, Worldtruthvideos, Thegreaterreset.org, Nordfront.dk, Lookaheadamerica.org, Patriotfront.us, Vastarinta.com and Nordicresistancemovement.org in Annex D.

[66] See Sections 4.1 and 4.2 of the Facebook, YouTube, WhatsApp, Instagram, TikTok, Twitter and VK profiles in Annex B.

[67] See Sections 4.1 and 4.2 of the Google Drive and Dropbox profiles in Annex B and of the Google Docs, MediaFire and Mega.nz profiles in Annex D.

[68] See Sections 4.1 and 4.2 of the Bitchute, Gettr (discretional) and SafeChat (discretional) profiles in Annex D.

[69] See Sections 4.1 and 4.2 of the Facebook, YouTube, Instagram, WhatsApp, TikTok, Twitter, Discord and VK in Annex B.

[70] See Sections 4.1 and 4.2 of the Google Drive and Dropbox profiles in Annex B and of the Google Docs, MediaFire and Mega.nz profiles in Annex D.

[71] See Sections 4.1 and 4.2 of the Bitchute, Gab, Gettr (discretional) and SafeChat (discretional) profiles in Annex D.

[72] See Sections 4.1 and 4.2 of the Telegram profile in Annex B and of the Element profile in Annex D.

[73] See Sections 4.1 and 4.2 of the Telegraph, Archive.org, Files.fm, Tlgur, Pixeldrain, Uploadgram, File.io, Gofile.io and Anonfiles profiles in Annex D.

[74] See Sections 4.1 and 4.2 of the Rumble.com, Patriots.win, Parler, Brandnewtube, 8kun, Redvoicemedia.com, WeGoSocial, 88msn.com, Doxbin, Wimkin, Mzwnews.com, Worldtruthvideos, Xephula, Thegreaterreset.org, Nordfront.dk, Lookaheadamerica.org, Patriotfront.us, Vastarinta.com and Nordicresistancemovement.org profiles in Annex D.

[75] In the "Our commitment" page of BitChute.com, it is stated that BitChute "is about Freedom of Expression". See https://support.bitchute.com/policy/our-commitment

[76] Rumble.com explains that its creation resulted from the recent rise of "cancel culture". See https://corp.rumble.com/our-story/

[77] On its homepage, Gab is described as a "social network that champions free speech, individual liberty and the free flow of information online". See https://gab.com

[78] On its homepage, Parler is described as the "premier global free speech platform". See https://parler.com

[79] Gettr describes itself as a "brand new social media platform founded on the principles of free speech, independent thought and rejecting political censorship and "cancel culture." See https://gettr.com/onboarding

[80] In the words of Redvoicemedia, "We support FREE SPEECH and will not quit telling the truth just because big tech tyrants don't like what we say". See https://www.redvoicemedia.com/about/

[81] According to Wimkin, "Wimkin stands for World Must Know (WMKN) due to our stance on no fact checking and free speech". See https://wimkinhelp.com

[82] Xephula explains that "The XEPHULA Network was created in efforts to combat big tech censorship and offer a balanced social media experience to its users". See https://xephula.com/static/about

[83] See for example "Our path", available at https://nordicresistancemovement.org/our-path/

[84] See subsections “The number of services issuing transparency reports expressly addressing TVEC is increasing”, “Transparency reports on TVEC are generally more detailed”, and “Some degree of convergence in transparency reports on TVEC can be seen” of Section 3 of this report.

[85] A zero-knowledge proof is a type of advanced mathematical construct in which it is possible to express a statement so that a *prover* can convince a *verifier* of the truthfulness of said statement without learning anything else other than the fact that the statement is true. See generally (Huqing and Zhixin, 2013[317]). When zero-knowledge proof protocols are relied upon to offer privacy-driven functionalities, they tend to be referred to as zero-knowledge privacy.

[86] See Section 5 of the Mega.nz profile in Annex D.

[87] See Section 5 of the Justpaste.it profile in Annex D.

[88] Available at https://www.canada.ca/en/canadian-heritage/campaigns/harmful-online-content.html

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

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

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

[92] See https://bills.parliament.uk/bills/3137

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

[94] See subsection “Disclosures by Chinese Platforms” in Section 3 of the Report.

[95] 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

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

[97] 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.

[98] Note that the Microsoft Services Agreement applies only to consumer use of Teams, and not to enterprise use.

[99] The Digital Safety Content Report therefore does not include the enterprise use of Teams, for instance.