Organisation for Economic Co-operation and Development

DSTI/CDEP(2023)29/FINAL

Unclassified

English - Or. English

24 June 2024

DIRECTORATE FOR SCIENCE, TECHNOLOGY AND INNOVATION

COMMITTEE ON DIGITAL ECONOMY POLICY

Transparency Reporting on Terrorist and Violent Extremist Content Online, Fourth Edition

JT03546717

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.

Foreword

This is the fourth instalment in a series of 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 three reports were Current Approaches to Terrorist and Violent Extremist Content among the Global Top-50 Online Content-Sharing Services (OECD, 2020); Transparency Reporting on Terrorist and Violent Extremist Content Online: An update on the global top-50 content sharing services (OECD, 2021); and Transparency Reporting on Terrorist and Violent Extremist Content Online (OECD, 2022). Like the third edition, this report 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 Nora Beauvais under the guidance of Jeremy West. It incorporates feedback from delegates on earlier drafts, as well as feedback from the companies profiled in Annexes B and D. The report was approved and declassified by the Digital Policy Committee on 5 April 2024 and prepared for publication by the OECD Secretariat.

Note to Delegations:

This document is also available on ilibrary as:

OECD (2024), "Transparency reporting on terrorist and violent extremist content online: Fourth edition", *OECD Digital Economy Papers*, No. 367, OECD Publishing, Paris, https://doi.org/10.1787/901cb8cf-en.

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.

The use of this work, whether digital or print, is governed by the Terms and Conditions to be found at http://www.oecd.org/termsandconditions.

© OECD 2024

Executive summary

This is the OECD's fourth benchmarking report examining the policies and procedures related to terrorist and violent extremist content (TVEC) online, with a focus on transparency reporting, of the world's top 50 most popular online content-sharing services (the "popular services"). Like the third edition, this report also covers the 50 online content-sharing services that terrorist and violent extremist groups and their supporters exploit or rely upon the most (the "intensive services"). The first three reports provided a benchmark against which this fourth report assesses relevant developments.

Terrorist and violent extremist actors continually adapt their methods to technological developments. As governments and online platforms increasingly take measures to curb the dissemination of TVEC, terrorists and violent extremists make adjustments to avoid content moderation. On mainstream online platforms, for example, they have been developing tactics to evade automated detection tools. Meanwhile, sustained efforts by large platforms to combat TVEC have also caused a "displacement effect" whereby terrorists and violent extremists turn to alternatives (e.g. cloud platform websites, decentralised web technology, niche alt-platforms, and terrorist-operated websites).

Transparency reporting on TVEC online is crucial to assess the evolution and magnitude of the threat, evaluate the effectiveness and efficiency of online platforms' policies and actions to tackle this problem, as well as their impact on human rights, and build an evidence base to support policymaking and regulatory frameworks.

The key findings of this report are:

1. The popular and intensive services are more diverse, both ideologically and geographically.

The TVEC landscape is multi-faceted, encompassing a wide range of ideologies, from terrorist groups to violent extremist political movements and lone actors, and it is spreading across different types of content-sharing services and geographical regions. For the first time in this report series, the popular services' list includes a gaming service. This is noteworthy because gaming services are increasingly used by terrorist and violent extremist actors. In addition, three Indian platforms have joined this ranking. As for the intensive services' list, it features a self-proclaimed anarchist website for the first time and covers a wider spectrum of geographic regions and languages.

2. Overlap between the popular and intensive services remains low, highlighting the need to look at the TVEC landscape more comprehensively.

Only ten services appear on both the popular and intensive lists, compared to 11 in the third benchmarking report. However, many policy discussions and responses still tend to focus on the largest platforms. Paired with the finding that the intensive services tend to be less transparent than the popular services (see below), the takeaway is that neglecting smaller but intensive services risks under-scrutinising or even turning a blind eye to a core part of the problem.

3. The evidence shows mixed results regarding the clarity of the popular services popular services' definitions of TVEC, while most of the intensive services' still do not define or even expressly prohibit TVEC.

On the one hand, the definitions related to TVEC in the popular services' policies and procedures are, overall, clearer than in the previous report. Services are using more comprehensive descriptions of TVEC and related concepts, but new gaps among the services' approaches have emerged, with a proportion of them still using vague terminology (18%) or having become less precise. On the other hand, 60% of the intensive services still do not define or explicitly prohibit TVEC, or they simply have not established any governing documents.

4. Transparency reporting on TVEC reveals new gaps among popular services and remains rare among intensive services.

Seventeen of the popular services now issue transparency reports with specific information on TVEC, as compared to just five in the first edition, 11 in the second, and 15 in the third of this series. This represents the slowest year-to-year growth rate to date. For the first time in the series, one of the services (present on both the popular and the intensive services lists) that previously issued transparency reports with TVEC-specific information ceased this practice. In addition, three of the four newest Services to issue transparency reports on TVEC provide very limited information, both quantitatively and qualitatively. Furthermore, there is still significant heterogeneity among the popular services' reporting approaches, which continues to make data aggregation and cross-platform comparisons difficult, if not impossible.

Among the intensive services, only six issue transparency reports on their policies and actions concerning TVEC, against 8 previously, and the vast majority (5 of 6) also appear in the popular services list. The scarcity in transparency reporting on TVEC among the intensive services may be explained by the fact that many of them are operated by terrorist and violent extremist groups and supporters, or by free speech "absolutists" who deliberately let TVEC flourish on their platforms.

5. Content moderation approaches continue to pose risks for privacy, freedom of expression and due process.

Continuing a trend that began during the COVID-19 pandemic, popular services rely more heavily on automated tools to detect and remove TVEC, which has generally increased the removal of lawful content and unjustified censorship. Furthermore, half of the intensive services remain opaque regarding their approaches to content moderation; and most of them either have no notifications and appeal mechanisms in place, or do not provide any information in this regard. This raises questions regarding their efforts to ensure the respect of privacy, freedom of expression and due process.

6. New online safety laws and regulations are creating an increasingly fragmented transparency reporting landscape.

As new online safety laws and regulations come into force, content-sharing services are facing new obligations to issue transparency reports in multiple jurisdictions, and they face different reporting requirements in each of them.

To conclude, this report highlights the need for more precision in the Services' governing documents; more consistency in the metrics and methodologies used to prepare transparency reports; more transparency in their content moderation approaches; and more efforts to ensure due process and to safeguard human rights and fundamental freedoms.

Table of contents

Foreword 2

Executive summary 3

1 Introduction 7
In a rapidly evolving TVEC landscape, improvements in transparency reporting are needed 9
The evolution of the OECD's benchmarking series on TVEC 11

2 Scope, methodology and research design 13

3 Updated commonalities, developments and trends in the popular services' approaches to TVEC 17
Changes in the popular services list since the third benchmarking report 17
Popular Services are using more precise descriptions of TVEC and related concepts, but new gaps among their approaches have emerged 17
The overall number of popular services issuing transparency reports expressly addressing TVEC continues to increase, although at a slower pace 22
Transparency reports on TVEC continue to improve, but they also reveal important disparities in quality and precision 23
Convergence in transparency reports on TVEC stagnates, while regulatory fragmentation intensifies 25
Notification and appeal mechanisms and processes 32
Disclosures by Chinese platforms 33

4 Updated commonalities, developments and trends in the intensive services' approaches to TVEC 35
Despite small progress made by some intensive services to improve their governing documents, the majority still does not explicitly prohibit TVEC 36
Half of the intensive services remain opaque regarding their approaches to content moderation 37
The majority of intensive services either have no notifications and appeal mechanisms in place or do not provide any information 38
Transparency reporting on TVEC among intensive services decreases 39

5 Enacted and emerging TVEC-related laws and regulations 41
Australia 41
Austria 44
Canada 45
European Union 45

France 47
Germany 48
Ireland 49
New Zealand 50
Korea 51
Türkiye 51
United Kingdom 52
United States 53

6 Conclusion 55

Annex A. Global Top 50 Most Popular Online Content-Sharing Services 56

Annex B. Profiles of the Top 50 Services 61

Annex C. The Global Top 50 TVEC-intensive Services 298

Annex D. Profiles of the Top 50 TVEC-intensive 300

Annex E. Definitions 374

References 375

Endnotes 398

TABLES

Table 3.1. Popular services' approaches to defining TVEC and related concepts 18
Table 3.2. Popular services that issue transparency reports with TVEC-specific information 22
Table 3.3. Convergence in six aspects of TVEC transparency reporting 26
Table 3.4. Regulatory-specific transparency reports issued by popular services 29
Table 3.5. Popular services' approaches to content moderation 30
Table 3.6. Popular services' approaches to notifications and appeals 32
Table 4.1. TVEC definitions in intensive services' governing documents 36
Table 4.2. Intensive services' content moderation approaches 37
Table 4.3. Intensive services' notifications and appeals mechanisms 38
Table 4.4. Intensive services that issue transparency reports with TVEC-specific information 39

1 Introduction

Digital transformation has radically transformed the everyday lives of billions of people, creating new public spaces and markets. Chatting, sharing content, live-streaming, and shopping online open new possibilities for people to communicate, work, consume, learn and create, participate in democracy and the economy, and enjoy their rights in the digital age (OECD, 2022). Internet-based technologies such as social media platforms, video, chat and messaging apps, streaming, file sharing, and gaming platforms offer tremendous possibilities for socio-economic progress and improved well-being. However, such technologies are also misused by terrorist and violent extremist groups to coordinate activities, spread propaganda, and glorify their atrocities.

Terrorists and violent extremists exploit online content-sharing services to conduct a wide range of actions, from incitement, radicalisation, and recruitment, to training, planning, sharing instructions and financing (UN CTED, 2021). In the context of the COVID-19 pandemic, which increased the amount of time spent online and speeded the adoption of digital technologies by several years (McKinsey & Company, 2020), these bad actors exploited this trend to further propagate their ideologies. Consequently, in 2021, the risk of online radicalisation increased, particularly in the case of right-wing terrorism (EUROPOL, 2021).

In addition, terrorist and violent extremist groups and individuals are continually adapting to take advantage of the online environment. Almost five years after the livestreamed terrorist attack in Christchurch, New Zealand, the online TVEC problem not only still exists but is rapidly evolving. On 14 May 2022, for example, a white supremacist opened fire in a supermarket in Buffalo, New York, killing ten people and injuring three. Part of the terrorist attack was livestreamed on Twitch, and although it was shut down by the service in under two minutes after being seen by 28 viewers or fewer, the video was reposted across platforms and seen by millions before being entirely removed. Copies of the attacker's diary and manifesto were shared mainly through smaller platforms (Ofcom, 2022). Quarterly figures published by major social media platforms show an increase in detected TVEC over the past years and research indicates that content from terrorist groups and their supporters – particularly the Islamic State – has surged on mainstream online platforms, since 2022 (Naffakh, 2022). To this day, videos glorifying the Christchurch terrorist attack are still easily discoverable on some of the most popular platforms, as well as content and account profiles supporting extremist ideologies like white supremacy and Holocaust denial (O'Connor & Smith, It is (still) shockingly easy to find terrorist content on TikTok, 2023).

Despite recent improvements in detection methods and renewed efforts from governments and online platforms to curb the dissemination of TVEC, terrorists and violent extremists are developing ever more creative ways to circumvent content moderation, even on larger platforms. For example, "jihadi" terrorist groups set up disinformation media outlets to whitewash official content from Islamic State group channels (Naffakh, 2022). Other techniques used to spread TVEC whilst evading content moderation tools, whether automated or manual, include:

- Modifying recognisable content, e.g., using alter-effects to hide parts of a video; using 'emojis' to replace certain words; favouring non-English languages; using video game footage to recreate an attack (Scott, Islamic State evolves 'emoji' tactics to peddle propaganda online, 2022);
- Using "broken text", which consists of breaking words up by punctuation or other symbols (Naffakh, 2022);

- Borderline content, which encompasses user-generated content that is either "borderline illegal content" (not technically illegal but with the potential to cause harm) or "borderline violative content" (not clearly violating a platform's policy) (GIFCT, 2023);
- "Trial and error" on new platforms, as terrorist and violent extremist actors regularly test and seize the potential of new platforms to spread TVEC (Naffakh, 2022);
- Abusing emerging technologies, e.g., using AR/VR technologies to facilitate remote training activities; or generative AI to generate high volumes of propaganda and create recruitment chatbots (Tech Against Terrorism, 2023).

Moreover, as new platforms continue to emerge and as more users can potentially be exposed to TVEC, terrorists and violent extremists are getting more adept at leveraging online content-sharing services to achieve their ends. Box 1 summarises four trends for terrorist and violent extremist use of the Internet, which were identified in a report by Tech Against Terrorism (Trends in Terrorist and Violent Extremist Use of the Internet | Q1-Q2 2021, 2021):

Box 1. Trends in terrorist and violent extremist use of the internet (Q1 – Q2 2021)

- **Increased use of "cloud platform" websites:** Islamist organisations including al-Qaeda, Islamic State (IS), and their supporter networks are increasingly exploiting open-source software to create "cloud platform" websites to store their content. These are password-protected websites that enable terrorist actors to share content via URLs. Many of them contain an extensive and regularly updated archive of terrorist material. This trend is likely due in part to a broad improvement in moderation of terrorist content by mainstream tech platforms. Cloud platforms currently provide terrorist actors with a comparatively stable, centralised location in which to store their material. This is because the process of taking down cloud platforms is extremely challenging.
- **Increased and diversified use of the decentralised web:** The exploitation of the decentralised web – or Dweb – by terrorist and violent extremist actors has both expanded and diversified. Messaging apps and social media platforms built on Dweb technology are serving critical roles in the online terrorist and violent extremist ecosystem, ensuring the ongoing availability of TVEC. Decentralised web hosting software and file storage systems like Skynet and the InterPlanetary File System (IPFS) are also increasingly being exploited for the hosting of terrorist content. The administrators of a prominent pro-IS propaganda archive website, for example, have been using a Dweb browser plugin since at least late 2020 to circumvent frequent takedowns. The plugin enables users to locate a stable landing page on which the latest link for the website can be found. This shift is likely the result of a combination of improved moderation by centralised platforms alongside a flawed perception among terrorist and violent extremist actors that Dweb services cannot be moderated.
- **Resurgence of terrorist operated websites (TOWs):** TOWs are websites that are run by terrorist actors and have been created for the sole purpose of furthering the goals of a terrorist organisation or network. This may be through the dissemination or archiving of terrorist content or recruitment of members. The resurgence of TOWs is likely a side-effect of broad improvements in social media platforms' content moderation efforts. As terrorist content moderation by mainstream platforms has strengthened, and the deplatforming of terrorist actors has become more widespread over the past few years, terrorist actors have been pushed onto increasingly niche platforms where the reach of their messaging is limited. As a result, terrorist actors and their supporters have increasingly supplemented accounts on smaller platforms with

their own sites and platforms. TOWs are often still indexed on search platforms and are often more easily discoverable in comparison to private channels on niche messaging apps.

- **Far-right extremist actors migrating to increasingly niche alt-platforms:** Violent far-right actors are migrating to increasingly niche alt-tech video-sharing platforms, as medium-sized platforms increase their capability to moderate and remove TVEC. Alt-tech platforms are often created in response to perceived notions of censorship on mainstream platforms. As alt-tech platforms champion themselves as advocates of "free speech" and regularly boast that they host content that has been removed elsewhere online, these spaces become havens for extremist actors seeking to evade the strict parameters of mainstream platforms.

Source: (Tech Against Terrorism, 2021)

In a rapidly evolving TVEC landscape, improvements in transparency reporting are needed

Although there is no internationally agreed definition of terrorism, most of the global top 50 most popular content-sharing services prohibit in their Terms of Service or Community Guidelines the use of their technologies to foster terrorist and/or violent extremist activities. They deploy a range of proactive and reactive measures to prevent, detect, and remove TVEC, including through automated means. In case of violation of their policies, they may take enforcement actions such as sending warnings, blocking or removing content, suspending and permanently banning accounts, or referring users to legal authorities (OECD, 2022).

However, new terrorist attacks regularly occur, showing the magnitude of the threat posed by online radicalisation and subsequent acts of violence in the offline world (Binder & Kenyon, 2022). The violent incidents of Christchurch (New Zealand), Poway (USA), El Paso (USA), Baerum (Norway), Halle (Germany), Buffalo (USA), and Bratislava (Slovakia) all had in common that the perpetrators were part of transnational online communities and took inspiration from one another (OECD, 2022). For some, their manifestos indicate the role of online propaganda in their radicalisation process (EUROPOL, 2022).

Besides, TVEC proliferates on a wide range of platforms, taking advantage of all kinds of features. For instance, the attacks in Christchurch and Buffalo highlight the growing phenomenon of gamification of violence, where extremist attackers repurpose elements from video games as part of their violence. Moreover, as livestream features become increasingly widespread across platforms, it also raises new challenges for content moderation. Livestreamed content is particularly difficult to moderate and enforce via Community Guidelines designed by platforms. First, it occurs in real-time and requires a fast and coordinated response. Secondly, automated tools have more difficulty identifying specific symbols or audio. Lastly, the content can be easily reposted to other forums after being removed from the original platform (Lamphere-Englund & White, 2022). The changing nature of the TVEC landscape, combined with the scarcity of transparency reporting on TVEC among content-sharing services, makes it difficult to assess the magnitude of the threat, its evolution, or the effectiveness of counter-measures.

On several occasions, governments and international fora have urged content-sharing services to take more drastic measures to curb the dissemination of TVEC online (G20, 2019; G7, 2019; G20, 2017; Christchurch Call , 2019). As explored in Section 5 of this report, many OECD countries have introduced bills and otherwise implemented initiatives to help address the problem of TVEC online. Among the most recent developments, the European Union adopted the Terrorist Content Regulation (TCO) with a view to stop terrorist propaganda and viral livestreams of violent attacks on social media. Having entered into force in 2022, it enables governments within the European Union to require online platforms to remove within an

hour specific posts, music, livestreams, photos, and videos inciting violence and glorifying terrorist attacks (Goujard, 2022).

Online content-sharing services have taken action to tackle TVEC on their platforms, notably by using a range of automated tools in their content moderation processes to curate, organise, filter, and classify potentially violating content. Fuelled by artificial intelligence and machine learning, such tools include for example digital hash technology, image recognition, metadata filtering, or natural language processing (NLP) (New America, s.d.). Furthermore, as discussed in Section 3 of the first and second benchmarking reports, Facebook, Microsoft, Twitter (now X), and YouTube formed the Global Internet Forum to Counter Terrorism (GIFCT) in 2017 with an aim to prevent terrorists and violent extremists from exploiting digital platforms (GIFCT, 2022), through research, technical collaboration and knowledge sharing. In addition, the GIFCT launched the Hash-Sharing Consortium, through which members could access a database of "hashes" or "digital fingerprints" of known TVEC that had been already detected and removed by at least one participating company. However, in 2022, the GIFCT restricted access to the hash-sharing database to GIFCT members only (GIFCT, 2022). Empowered by this database and in accordance with their internal policies, GIFCT 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, 2022). Other initiatives include, for example, the Hash-Matcher-Actioner (HMA), launched by Meta in 2022, an open-source software tool designed to help platforms identify copies of images or videos and take action against them (Clegg, 2022).

At the same time, this has raised concerns about both the capacity of online services to efficiently counter TVEC on their platforms and the risks related to intrusive monitoring of content, 'false positives', and the potential negative impacts on individuals' human rights and fundamental freedoms, such as freedom of expression and the right to privacy as outlined in Article 17 of the International Covenant on Civil and Political Rights (ICCPR) (OECD, 2022).

Within this context, transparency and accountability for content moderation practices can be effective in addressing both of those concerns. The Santa Clara Principles on Transparency and Accountability in Content Moderation lay out essential transparency practices that companies could adopt to guarantee the protection of human rights and due process. Among other things, they exhort companies to publish understandable rules and policies, to ensure that content moderation decisions take into consideration the diversity of cultures and contexts in which the services operate, to apply principles of accuracy and non-discrimination in detection methods, and to make notice and appeal mechanisms available to users. Since their launch in 2018 and actualisation in 2021, twelve major companies have endorsed these principles and the overall number of companies providing transparency and procedural safeguards has increased (Santa Clara University's High Tech Law Institute, 2021).

Although there appears to be a trend towards greater transparency and granularity from companies (Radsch, 2023), even among many of the global top 50 most popular services that issue regular transparency reports, significant progress needs to be made before they reach the level of detail and quality necessary for readers to derive meaningful insights, assess their efficacy and impacts on human rights, and get a cross-industry perspective. Sections 2 and 3 of this report reviews commonalities and differences among the services' approaches and highlights convergence in key aspects of TVEC transparency reporting. New regulations such as the Digital Services Act in the European Union, could, to some extent, pave the way for more systematic, coordinated transparency reporting practices, in a way that respects human rights (Article 19, 2023).

The presence of TVEC online has proven to be in some cases a catalyst for the radicalisation of individuals to violence and the commission of terrorist acts, with severely negative impacts for users, citizens, and society at large. There are also important adverse consequences for online services as TVEC undermines the trust of their users and damages their business models (European Union, 2021). Enhanced

transparency reporting on TVEC would enable better assessments of the threat through more reliable, consistent, and comparable data, ultimately improving evidence-based policymaking and regulatory frameworks for a safer online environment.

The evolution of the OECD's benchmarking series 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 periodic 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 (OECD, 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. It identified commonalities, developments and trends in their approaches. An important aspect of this inquiry was whether the popular 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, an Update on the Global Top 50 Content-Sharing Services (OECD, 2021), focused on the degree to which the popular services' approaches to tackling TVEC online had changed and evolved over the course of one year. This report specified whether there was more or less clarity in how the popular 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.

The third benchmarking report (OECD, 2022) was broader in scope, as it took into account the "displacement effect" resulting from more aggressive content moderation by a number of large online platforms. The report showed that stronger responses by large platforms to address TVEC resulted in switching patterns on the part of terrorists and violent extremists. These bad actors turned to small- or micro platforms lacking either the financial and technical resources to make their services unappealing to them, or the will to do so (Donovan, Lewis, & Friedberg, 2019). Therefore, in addition to providing an update on the global top 50 Services' approaches to countering TVEC online, the third edition explored the TVEC-related policies and procedures of the 50 Services that terrorist and violent extremist organisations and individuals rely upon the most, the intensive services.

The present report is comprised of two main parts. The first explores the degree to which the popular services' approaches to TVEC online have evolved since the publication of the third benchmarking report. The second does the same with the Intensive Services[1], highlighting commonalities, variances, and trends among them. As with the preceding three benchmarking reports, this report provides an objective and factual snapshot in time. It also informs a project led by the OECD, in consultation with experts from government, business, civil society and academia, that has led to a multi-stakeholder, consensus-driven framework for voluntary transparency reporting on TVEC online by content-sharing services. The framework contains a standardised template that all companies wishing to report on TVEC can use, and that all OECD members support.

The report is structured as follows. Section 2 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 selection criteria for the intensive services. Section 3 summarises the first three benchmarking reports' main findings and explores the development and evolution of the popular services' approaches to tackling TVEC online

since the last edition. Section 4 summarises the third benchmarking report's main findings with respect to the intensive services' TVEC-related policies and procedures, then identifies variances, similarities and trends based on the updated list and profiles in this report. Section 5 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 third benchmarking report.

2 Scope, methodology and research design

The main objective of the first benchmarking report was to determine the state of play among the popular services with regard to their policies, procedures and practices relevant to TVEC. The popular 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 popular services were chosen on the basis of the size of their user bases – i.e. the extent to which they are "popular" – under the assumption that TVEC on these Services would reach a wider audience, which is a known criterion of terrorist groups when choosing their platforms (Tech Against Terrorism, 2019). One year later, the second benchmarking report followed the same methodology and tracked the evolution of the popular services' approaches to tackling TVEC online over that one-year period.

Recognising the impact of the displacement effect on the dissemination of TVEC online, which caused terrorist and violent extremist groups to migrate from mainstream to smaller, security-oriented and fringe platforms, the third benchmarking report broadened the scope of the platforms considered. With a view to examining services on which the most TVEC actually appears, rather than focusing solely on the most popular services, it looked at the approaches to TVEC of services grouped according to two distinct rankings: the top 50 popular services and the top 50 intensive services.

Building on the methodology used in the third edition, this report focuses on the approaches to TVEC of both the Popular and intensive services, assessing the evolution of online content-sharing services' approaches to TVEC since the previous edition. Specifically, it looks at 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 the Services follow to detect and address such content are clear and transparent, and whether the number of Services that publish transparency reports on TVEC has changed.

As in the first three 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, online communications services, and gaming 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 popular services is included in Annex A. The TVEC-relevant policies, procedures and practices of the top 50 Services are presented in Section 3, highlighting commonalities, variances and trends in them.

A list of the world's top 50 intensive services is included in Annex C. Likewise, the TVEC-relevant policies, procedures and practices of the TVEC-intensive Services are presented in Section 4, highlighting commonalities, variances and trends in them.

In the third benchmarking report, to generate the list of the top 50 TVEC-intensive Services in a manner that best depicted 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. The methodology used aimed to capture, on one hand, the different ways in and purposes for which terrorist and violent extremist groups rely on digital technologies, and on the other hand, the likely reach, or success, of such groups' TVEC dissemination practices. On the basis of information gathered on "jihadi" and far-right groups for the period January-December 2021, SITE divided the Intensive Services into three categories:

- Mainstream TVEC-intensive Services: Services 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.
- File-sharing TVEC-intensive Services: Services 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.
- Far-right-focused TVEC-intensive Services: Services either created by, predominantly exploited by, or accommodating to far-right extremists. This category did not include "jihadi" sites. Accordingly, the platforms and websites in this category were 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.

For this edition, to identify the top 50 TVEC-intensive Services, the OECD partnered with the Terrorism Research & Analysis Consortium (TRAC), a global consortium of 2,800 experts who live in and report from around the world, combining one of the world's largest databases of terrorists, terrorist groups, violent hate groups and their abettors with original, analytical essays on terrorism, profiles of vulnerable regions and cities, and a live feed of news and analyses.

Differing from SITE's quantitative approach, TRAC's methodology relies more on qualitative research techniques, and notably netnography, an adaptation of traditional ethnography which explores social interactions, behaviours, and perceptions within online communities. Netnography's unobtrusive nature enables analysts to observe discussions, social rituals, and TVEC sharing methods, providing insights into

both immediate behaviours and the broader TVEC landscape, including underlying ideologies, motivations, and experiential perspectives, which are all essential components to contextualise TVEC.

The TVEC landscape is characterised by rapidly changing terminologies and sudden group migrations from one service to another. TRAC drew on its pre-existing access to terrorist and violent extremist gated online communities to collect data. The netnographic approach developed by TRAC was rooted in continuous observation, offering an immersive understanding of the situational context, real-time interactions, and the processes of posting, crowd-sourced editing, and occasional discussions following the swift removal of TVEC. TRAC analysts were able to scrutinise the ongoing and repetitive interactions among terrorists and violent extremists in various social media channels. They analysed their communications with one another, their engagements with affiliated or unaffiliated groups, and how they created and disseminated TVEC.

While this method is time-consuming, it offers the following advantages:

- The collected data comes directly from the source, not from proxies or external observations;
- TRAC was not only able to monitor current TVEC storage locations, but also to gain insight into the dissemination strategies employed;
- The spectrum of TVEC considered is more extensive than in the previous report, both ideologically and geographically;
- This approach acknowledges the importance of human oversight to understand and analyse the complexity of the TVEC ecosystem, in contrast to a strictly quantitative approach.

TRAC selected the most recent, pertinent, and influential TVEC dissemination services to date based on its observations and experience. Some of them, such as Telegram, serve as hubs for the most virulent and prolific TVEC originators. The services are no longer divided into categories, as was the case in the third benchmarking report. Instead, TRAC ranked them by order of importance, devoting 12 spots to file-sharing platforms (profiles 34 to 45), and 5 spots to encrypted communication platforms (profiles 45 to 50). It should be noted that, despite the methodological changes, 22 of the Services appearing on this edition's list substantially overlap with those on the previous edition's list also appeared in the list prepared by SITE in the third benchmarking report, and they still include a majority of mainstream and file-sharing platforms used by "jihadi" and far-right groups.

The review of the popular services' and the intensive services' approaches to combatting TVEC online follows three main steps.

- First, a standardised template is used to profile each Popular Service and Intensive Service, based on the Service's publicly available terms of service (ToS), community guidelines and policies, blogs, service agreements and other official information ("governing documents")3. All services are contacted and asked to provide feedback on the accuracy of their profiles and any additional relevant information.
- Second, all profiles are updated based on the services' responses. These updated profiles are included in Annexes B and D.
- Third, the developments and changes in the Popular and intensive services' approaches to TVEC since the third benchmarking report, are identified and summarised.

Key aspects in the popular services' and 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.

3 Updated commonalities, developments and trends in the popular services' approaches to TVEC

Changes in the popular services list since the third benchmarking report

There are two noteworthy changes in this list as compared to that in the third benchmarking report. First, this edition includes a gaming service (Steam) for the first time in the report series. This is noteworthy because gaming services and associated platforms are increasingly used by terrorists and extremists to disseminate digital propaganda and for purposes of radicalisation and recruitment (Lakhani, Video Gaming and (Violent) Extremism - An exploration of the current landscape, trends, and threats, 2021). They are also often targeted at younger users. According to the Global Network on Extremism & Technology, potential threats include gamification for radicalisation, exploiting gaming as pop culture to appeal to younger generations, exploiting online games for communication, financing, and money laundering (Lamphere-Englund & White, 2023). During the COVID-19 pandemic, the combination of social distancing and lockdown measures exacerbated the risks posed by TVEC online, particularly among younger people and minors (EUROPOL, 2022).

Second, three Indian platforms have joined the ranking of Services. This can be partly attributed to the fact that, since 2020, the Indian government has issued several bans on Chinese platforms and apps, which has driven users to Indian platforms and apps such as Josh, an alternative to TikTok (Hemjrajani, 2022). Under the Indian Information Technology Act, the Indian central government can block access to any domain or app that is deemed to be a threat to national security. Successive bans have resulted in hundreds of Chinese services becoming unavailable in India, which has in turn created demand for local alternatives.

Popular Services are using more precise descriptions of TVEC and related concepts, but new gaps among their approaches have emerged

The first two benchmarking reports found dissimilar approaches in the popular services' definitions of TVEC and related concepts, with few providing definitions that allow a clear understanding of what type of content is prohibited and what is considered a terrorist and/or extremist group or organisation. However, the third benchmarking report identified a trend towards more definitional clarity with more popular services providing comprehensive definitions and examples.

This report finds a continuation of that trend, with more popular services having improved their definitions of TVEC and related concepts and provided additional explanations and examples to clarify what is and is not allowed on their platforms. The figures in Table 3.1 show that, overall, about one third of the popular services explicitly ban the use of their technologies to foster terrorism and/or violent extremism and provide sufficient detail to understand the scope of their policies. Only nine popular services use broad and/or vague descriptions of prohibited conduct, which descriptions can be interpreted as supersets encompassing TVEC, although without specifically prohibiting it. In spite of the overall progress, though, a minority of popular services changed their terms of service and/or community guidelines in ways that make them less clear, less precise, or less restrictive than before.

Some of the popular 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 instance, YouTube updated the categories of violating content featured in its Community Guidelines. The former "Violent Criminal Organisations policy", under which TVEC fell, was changed to the "Violent Extremist and Criminal Organisations policy", which places a new focus on violent extremist organisations. It also provides additional examples of prohibited material, such as content that glorifies or promotes violent tragedies (e.g., school shootings) (YouTube/Google, 2023).

Table 3.1. Popular services' approaches to defining TVEC and related concepts

	1st benchmarking report	2nd benchmarking report	3rd benchmarking report	4th benchmarking report
Services that explicitly ban the use of their technologies to foster terrorism and/or violent extremism, and that define terrorism and/or violent extremism and related concepts with sufficient detail to understand their scope, providing examples where appropriate	5[4]	6[5]	11[6]	18[7]
Services that explicitly ban the use of their technologies to foster terrorism and/or violent extremism, using but not explaining in detail such terms and related concepts	19[8]	21[9]	19[10]	21[11]
Services that include TVEC within the same category as hate speech, hateful content and/or violent or graphic content	15[12]	13[13]	5[14]	2[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]	9[19]

Facebook and Instagram (both owned by Meta) refined their "Dangerous Individuals and Organisations" policy with more examples of prohibited behaviours (e.g. praise, substantive support, or representation of the perpetrator(s) of such attacks; perpetrator-generated content relating to such attacks; or third-party imagery depicting the moment of such attacks on visible victims) (Facebook, 2023). Furthermore, Facebook and Instagram replaced the terminology "hate organisations" by "hate entity", which is defined as an organisation or individual that spreads and encourages hate against others based on their protected characteristics. This new policy extends to individuals (instead of just organisations) and provides examples of activities that characterise a hate entity, such as violence, threatening rhetoric, or dangerous forms of harassment targeting people based on their protected characteristics, repeated use of hate speech, representation of hate ideologies or other designated hate entities, and/or glorification or substantive support of other designated hate entities or hate ideologies.

Likewise, LinkedIn updated its Professional Community Policies and improved its categories of violating content with more detailed explanations. TVEC is now included in a new category titled "Dangerous

Organisations and Individuals"; whereas previously, LinkedIn simply required users not to post terrorist content or promote terrorism and violent extremism. In substance, the new policy prohibits the same content and behaviours, but also extends to other types of dangerous organisations and provides a list of examples of what constitutes TVEC. It prohibits users from posting content that visually depicts acts of terrorism or that promotes or propagandises terrorist groups, individuals, or activities; the glorification or incitement of dangerous organisations and individuals, acts of terror or violent extremism; recruitment for terrorist groups or other organisations that espouse violence; as well as profiles or pages created and maintained by or in support of terrorist organisations or terrorist individuals. In addition, LinkedIn clarifies that terrorist organisations and terrorist individuals include "any non-state group that (1) identifies through its stated purpose, publications, or actions as an extremist group, (2) engages or has engaged in violence and/or the promotion of violence to further its cause, and (3) targets civilians (non-military)"; as well as "any individual that appears in the US Federal Bureau of Investigation's Domestic Terrorism or Most Wanted Terrorists Lists or is a Specially Designated Global Terrorist (SDGT) by the United States Department of State or the US Department of the Treasury" (LinkedIn, 2022).

In 2023, Snapchat published an "explainer" with detailed information on its policy regarding "Hateful Content, Terrorism, and Violent Extremism" (Snapchat, 2023). It defines TVEC as "content that promotes terrorism or other violent, criminal acts committed by individuals or groups to further ideological goals". These rules also prohibit any content that promotes or supports foreign terrorist organisations or extremist hate groups – as designated by credible, third-party experts – as well as recruitment for such organisations or violent extremist activities. Moreover, Snapchat published "Content Guidelines for Recommendation Eligibility" under which accounts that repeatedly or egregiously violate the eligibility criteria may be temporarily or permanently disqualified from algorithmic recommendations. Terrorism and violent extremism are expressly cited within the scope of the Guidelines (Snapchat, 2023).

Reddit's Content Policy now specifically defines terrorist content, whereas before it broadly prohibited threats of violence, as "propaganda material posted by terrorists or designated terrorist organisations and their supporters, expressions of affiliation or support for terrorists or designated terrorist organisations, glorification of terrorist acts or content that solicits or incites a person or group to participate, commit, or contribute to terrorist activities" (Reddit, 2023). Examples of prohibited content of relevance to TVEC are provided, such as posts or comments with a credible threat of violence against an individual or group of people, and posts containing mass killer manifestos or imagery of their violence. The specific prohibition of posting manifestos is important in light of the mass shooting that occurred in 2022 in Buffalo, New York, as the perpetrator had shared online[20] a white supremacist manifesto before the attack (Anti-Defamation League, 2022).

For its part, Twitch introduced in 2022 a new username policy which prohibits references to terrorism or terrorist organisations, threats, and hateful conduct in account usernames and display names (Twitch, 2022). Twitch registered a notable increase in terrorism enforcements (+403%) between H2 2021 and H1 2022, noting that 67% of these enforcements were for "Terrorism via Username", which indicates that the introduction of the new policy likely played a significant role in this increase. Furthermore, in H2 2022, Twitch updated its Community Guidelines with simplified language, additional context and examples of prohibited content under its policies (e.g., encouraging others to participate in acts that may harm others) (Twitch, 2023).

In addition to prohibiting the organisation, promotion or support of violent extremism, Discord now forbids the glorification of violent events or the perpetrators of violent acts, as well as promoting conspiracy theories that could encourage or incite violence against others. However, its Community Guidelines still do not specifically mention terrorism (Discord, 2023).

WeChat has substantively improved its definitions of TVEC and related concepts. Previously, WeChat broadly prohibited terrorist activity and violent acts, but its new policy titled "Terrorism, Violent Extremism

and Other Criminal Behaviour" provides definitions for both "terrorism" and "violent extremism", as well as much more granular information on what they encompass. Terrorism is defined as the "use of political violence and the exploitation of fear aimed at reaching at a target audience, with the end goal of bringing behavioural and societal change through coercion, whether that be to further a religious, political or any other ideological cause". Violent extremism is defined as "a form of extremism that condones and enacts violence with ideological or deliberate intent, such as religious or political violence." WeChat adds that violent extremist views can manifest in connection with a range of issues, including politics, religion, and gender relations; and that extremism is a term used to characterise a variety of attitudes, beliefs, and behaviours that often are on the extreme end of the political, religious, or ideological spectrum within society (e.g., white nationalist, anarchist) (WeChat, 2023). Lastly, WeChat underlines that it takes a very strict stance against terrorism and violent extremism and does not allow dangerous individuals or organisations to use the platform to promote terrorism, violent extremism, crime, or other types of harmful activities. As a result, it may suspend or terminate a user's account and notify the relevant legal authorities if it believes there is a threat to the safety of others (WeChat, 2023).

Among the popular services that appear in this report for the first time, Dailymotion and Josh both provide definitions for TVEC. Dailymotion defines terrorist content as "any content that advocates or promotes violent extremist and/or terrorist organisations, individuals, or acts" and relies on official consolidated lists of violent extremist organisations published by the United Nations and the European Union. Dailymotion gives examples of content that may be considered TVEC, including but not limited to content provoking the commission of terrorist offences or violent extremist acts or glorifying such offences; encouraging participation in terrorist offences; praising, glorifying, and/or supporting the acts of violent extremist and/or terrorist individuals or groups; supporting terrorist and/or violent extremist ideologies; encouraging people to join violent extremist and/or terrorist organisations; providing instructions on methods or techniques for the commission of terrorist offences; containing images, sounds or symbols (e.g., names, logos, flags, slogans, uniforms, gestures, pictures) intended to depict violent extremist and/or terrorist organisations or individuals; emanating from criminal, violent, extremist, or terrorist organisations; depicting one or more hostages; or posted online with the aim of soliciting, threatening, or intimidating people on behalf of a criminal, violent, extremist, or terrorist organisation.

For its part, Josh uses a much broader definition of "terrorist and extremist content", as "any content which threatens the security and national integrity of India or promotes any dangerous activities are a violation of Josh's Platform's Policies". Josh adds that users must not use the platform to incite terrorism, secession, and acts of violence, or post any content that promotes or encourages users to take actions on behalf of a terrorist or anti-national organisation, or recruits and disseminates information for, or furthers the objectives of, such organisations. Lastly, users must not garner support or approval for the commission of violent acts during lawful protests or create violence-inducing terror and conspiracy networks on the platform.

Another trend is that an increasing number of popular services expressly refer to designation lists in their Terms of Service and/or Community Guidelines. In the third benchmarking report, that was already the case for Facebook, Facebook Messenger, Instagram, Quora, Microsoft's Services (Skype, Teams, LinkedIn, OneDrive), Vimeo, Wordpress.com, and YouTube. In addition to these, TikTok's definition of "violent extremists" now references groups designated as such by the United Nations. As mentioned above, Dailymotion relies on official consolidated lists of violent extremist organisations published by the United Nations and the European Union; while Snapchat considers designations by credible, third-party experts, although it does not specify which ones. Lastly, when defining "hate symbols", Snapchat refers inter alia to the hate symbols database maintained by the Anti-Defamation League (Anti-Defamation League, 2023). On the contrary, X's updated policy does not mention "national and international designations" as was previously the case.

A minority of popular services that used to provide a certain degree of precision have implemented changes to their policies resulting in less clarity or precision. First, TikTok replaced its policy on "Violent Extremist

Organisations and Individuals" with a new category titled "Violent and Hateful Organisations and Individuals" (TikTok, 2023). Under this policy, terrorist organisations are neither mentioned nor defined, whereas before they were clearly defined and prohibited. However, TikTok defines violent extremists as "non-state groups, including those designated by the United Nations, that threaten or use violence against civilians for political, religious, ethnic, or ideological reasons". Second, X (formerly Twitter) changed its "Safety" policies so that they no longer feature the category "Terrorism and Violent Extremism" (X, 2023). As a result, TVEC falls under a broader category titled "Violent Hateful Entities", which encompasses terrorist organisations and violent extremist groups but also perpetrators of violent attacks and individuals who affiliate with and promote their illicit activities. Additionally, this policy appears to be broader in scope, as it allows a number of exceptions for violent and hateful entities. Such exceptions can apply, for example, if they have reformed or denounced their violence and/or hate-based purpose; if they are engaged in a peaceful resolution process; or if they are a state or governmental entity, including if they have representatives elected to public office. In addition, in Google Drive's updated Abuse Program Policies, TVEC now falls under the "Violent Organisations and Movements" category, whereas before it was in a standalone category dedicated to "terroristaActivities". In substance, the same content and behaviours are prohibited but the scope of this category extends beyond just TVEC. Likewise, IMO updated its Community Guidelines, which do not feature a specific category for "Terrorism and Violent Extremism" any longer. Presently, TVEC is conflated in two larger categories ("Politics / Religion / Terrorism" and "Violent Extremism / Crime Organising").

Finally, some popular services have implemented changes to their enforcement policies and sanctions in response to TVEC, and adopted a stricter approach. For instance, YouTube's strike system now provides additional penalties for channels that have received a strike. Under this system, after one strike, users may not start a scheduled live stream or schedule a video to become public for one week. Furthermore, a single case of severe abuse can now sometimes result in channel termination without warning (YouTube/Google, 2023). In the same vein, Snapchat established a strike system that creates a record of violations for users. When hateful content is reported, Snapchat will remove any violating content and users who engage in repeated or egregious violations will have their account access locked. Accounts that are used primarily to violate the Community Guidelines or to perpetrate serious harms, such as promoting terrorist and extremist activity, are immediately disabled (Snapchat, 2023). Likewise, in the case of TVEC, LinkedIn may permanently restrict an account after a single violation. And Pinterest may limit or remove boards or accounts for repeated violations or after a single instance of a severe policy violation (Pinterest, 2023).

The three first benchmarking reports found that, when defining and identifying a terrorist/violent extremist organisation, the popular services had different approaches. Over the last year, disparities among the popular services' approaches have widened between, on the one hand, those that are using more comprehensive and precise definitions, and on the other hand, a minority that either was already using vague and/or broad terms or has lost clarity.

When defining and identifying TVEC and terrorist and/or violent extremist organisations, the same general types of approaches found in the third benchmarking report remain. First, a few popular services including Facebook, Instagram, Snapchat, WeChat, and Dailymotion provide their own definitions, breaking down this concept into different sub-categories such as hate organisations, violent non-state actors or "other" violent organisations[21]. Second, a growing share of popular services, such as LinkedIn, TikTok, Vimeo, and YouTube, rely on government or United Nations lists of terrorist organisations. Lastly, the remaining popular services are still silent in this regard.

The overall number of popular services issuing transparency reports expressly addressing TVEC continues to increase, although at a slower pace

The number of popular services issuing transparency reports with specific information on TVEC continues to increase. In the first benchmarking report, this number was five. Three years later, it is 17. (See Table 3.2) However, three of the four newest popular services to issue transparency reports containing information specifically about TVEC provide very limited information, both quantitatively and qualitatively.

Table 3.2. Popular services that issue transparency reports with TVEC-specific information

1st benchmarking report	2nd benchmarking report	3rd benchmarking report	4th benchmarking report
5	11	15	17
Facebook	Facebook	Facebook	Facebook
YouTube	YouTube	YouTube	YouTube
Instagram	Instagram	Instagram	Instagram
Twitter	Twitter	Twitter	Wordpress.com
Wordpress.com	Wordpress.com	Wordpress.com	Skype
	Skype	Skype	OneDrive
	OneDrive	OneDrive	Twitch
	Twitch	Twitch	TikTok
	TikTok	TikTok	Reddit
	Reddit	Reddit	Discord
	Discord	Discord	Zoom
		Zoom	Snap
		Snap	Teams
		Pinterest	Dailymotion
		Teams	Moj
			ShareChat
			Josh

The results should also be put in perspective with the fact that this edition shows the slowest rate of increase since the start of the report series. Also, the new popular services that publish transparency reports with TVEC-specific information (namely Dailymotion, Moj, ShareChat, and Josh) do so with a variable level of precision, and all of them are new to the global top 50 popular services list. This means that no remaining Service from the previous reports started issuing transparency reports with information on TVEC for the first time.

Nonetheless, this edition does reveal a “first”. For the first time in this report series, one of the popular services that previously issued transparency reports with TVEC-specific information ceased this practice. X had been publishing regular transparency reports expressly addressing TVEC since the first benchmarking report. In 2023, Twitter was acquired by X Corp. and changed its name to X. In parallel, X is currently undergoing internal reorganisation and has stopped publishing transparency reports. The last one available is from 2022 and covers July to December 2021. Yet, according to researchers focused on online extremism, terrorist activity on X has increased by at least 69% since its acquisition by X Corp (Naffakh, 2022). In 2022, X announced the creation of a content moderation council with “widely diverse viewpoints”, explaining that no major content decisions or account reinstatements would happen before that council convenes (Frenkel, 2022). However, no new information has been released since then. X also announced that Elon Musk, X Corp.’s CEO, and Linda Yaccarino, Chief Executive, would both oversee the trust and safety team (Dang, 2023). It is not clear yet what these changes will entail for transparency reporting and content moderation at X in the future.

Nevertheless, the fact that the number of popular services issuing transparency reports with at least some specific information on TVEC has more than tripled after four years is noteworthy. In fact, the number of popular services issuing transparency reports on TVEC could be slightly higher. Telegram does not publish official transparency reports. However, it continues to disclose the number of ISIS terrorist bots and channels banned on a daily basis, and the aggregate monthly number, on its ISIS Watch channel (Telegram, s.d.).

Additionally, some popular services have started issuing transparency reports that do not break out specific information on TVEC but do include it in broader reporting categories. For instance, Baidu, the parent company of Baidu Tieba and iQIYI, has been publishing annual Environmental, Social and Governance Reports since 2020. These reports contain a section on "Content governance" with information on Baidu's content moderation across all its services. Although they do not contain TVEC-specific metrics, "Terrorism / Incitement to violence" is identified as an area of focus for content governance at Baidu (Baidu, 2022). Moreover, Kuaishou International Business (KSIB), the parent company of Kuaishou/Kwai, has started issuing semi-annual transparency reports in H1 2021, accounting for both Kwai (outside of the People's Republic of China) and Snack Video (another one of its short video apps). It contains information on videos removed broken down by removal reasons under which TVEC may fall (e.g., Dangerous Individuals and Organisations, Hateful Behaviour, Violent and Graphic Content); as well as on account removals and videos reinstated after a successful appeal (Kuhaishou/Kwai, 2022). Given the narrow scope of this information and the fact that the reporting categories are broader than just TVEC, Telegram, Baidu Tieba, and Kuaishou/Kwai have been left outside the group of popular services that issue transparency reports on TVEC.

The slow rate of increase in transparency reporting on TVEC among the popular services could be attributed to the following: first, producing a first transparency report presents difficulties, especially for small or new companies as it 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; second, the production of transparency reports is also determined by whether a given service has relevant policies and content moderation in place (OECD, 2022); and, third, for the first time in this benchmarking series, one Service has stopped issuing transparency reports on TVEC.

Transparency reports on TVEC continue to improve, but they also reveal important disparities in quality and precision

The third benchmarking report highlighted two main trends regarding the TVEC-specific content of transparency reports. First, the popular services that already issued transparency reports tended to include more granular information or new metrics in subsequent releases. This trend continues in this edition to a certain extent, with the exception of X. Second, the third report found that some of the popular services that already issued transparency reports, but without TVEC-specific information, supplemented subsequent reports with information about their efforts to tackle TVEC and/or explanations of their methodologies. This year, the second trend has stopped, as all the new transparency reports with TVEC-specific information come from popular services that are featured in this report series for the first time.

Since the third benchmarking report, Twitch started reporting on three additional metrics: the number and percentage of appeals received, broken down by category of violating content, including "Terrorism", "Violent Graphic Content", and "Hate/Harassment"; the number of appeals granted for each category; and the percentage of reports by response time (under 10 min, 30 min, 1 hour, 6 hours, 12 hours or 24 hours) – however, this last metric is not broken down by category of violating content. With regard to appeals, Twitch explains that the 12 successful ones for terrorism-related content were based on enforcement errors

for news or documentary footage that referenced but did not depict, glorify, encourage, or support terrorism or violent extremist acts or actors (Twitch, 2022).

TikTok also expanded the scope of information provided on TVEC. In addition to reporting the percentage of videos removed by removal reason (including "Violent Extremism", "Hateful Behaviour", and "Violent and Graphic Content"), as well as the corresponding removal rates (proactive removal rate, removal rate before any views, and removal within the 24 hours), TikTok provides the percentage of videos removed and corresponding removal rates by sub-policy. For the category "Violent Extremism", TikTok distinguishes between "violent extremist organisations and individuals" and "threats and incitement to violence"; for "Hateful behaviour" between "hateful ideology" and "attacks and slurs on the basis of protected attributes", and for "Violent and Graphic Content" no additional breakdown is provided. TikTok also started reporting on other metrics that are not TVEC-specific but that can be useful to better understand its policies and content moderation processes. First, TikTok communicates the human moderation language distribution (percentage of moderators assigned by language). Then, similarly to Twitch, TikTok also started publishing the response time to community-reported content (less than 2 hours, 2 to 8 hours, 8 to 24 hours, more than 24 hours) (TikTok, 2023).

For its part, Dailymotion released its first transparency report on tackling terrorist and violent extremist content in 2023, covering the period from January to December 2022, which is exclusively dedicated to TVEC. It discloses the number of TVEC removed following a detection by Dailymotion; the number and percentage of user reports received flagging potential TVEC; and the number and percentage of TVEC removed following a user report. Regarding appeals, it contains the number and outcome of complaints brought by a content provider concerning a content removal; the number and outcome of administrative or judicial review proceedings brought by Dailymotion; and the number of cases in which the content was reinstated following either one of these cases.

Some popular services made an effort to clarify their methodologies. Snapchat's latest transparency report features an updated glossary with additional context around the metrics used in the report and how they are calculated. It now defines "content and account reports", "enforcement", "total content enforced", "total unique accounts enforced", and "turnaround time" (Snapchat, 2023).

Facebook, Instagram, Discord, Wordpress.com, Snapchat, and Microsoft (Teams, Skype, OneDrive) continue to report the same metrics as last year.

Meta (owner of Facebook and Instagram) did do something new and noteworthy, though. It undertook an independent third-party assessment of its Community Standards Enforcement Reports for those two Services. The assessment was conducted by EY for the period 1 October 2021 to 31 December 2021 and concluded that the calculation of the metrics reported had been prepared based on the specified criteria, were fairly stated, and that Meta's internal controls were suitably designed and operating effectively (Sarang, 2022).

Among the new popular services that publish transparency reports with information on TVEC, the three Indian Services (ShareChat, Moj, and Josh) provide limited information. ShareChat and Moj, both owned by Mohallah Tech, use the same reporting template. Their transparency reports comprise only one TVEC-specific metric, which is the number and percentage of content and accounts reported, broken down by category of violating content, among which are "Terrorism", "Hate speech", "Illegal activities", and "Violence". Other information is included on law enforcement requests, user reports, takedowns, and bans, but not specifically for TVEC (Moj, 2023) (ShareChat, 2023). As for Josh, it published its first transparency report covering January to September 2022, showing the percentage of enforcement actions taken on content, broken down by category of violating content (as tagged by users), among which are "Terrorism and Extremism", "Violent and Graphic Content", and "Illegal Activities and Regulated Goods" (Josh, 2022). No number is provided, though, which makes it hard to get a sense of the actual volume of TVEC actioned by Josh.

Finally, another notable development is that some of the popular services already issuing transparency reports with information on TVEC do so at more frequent intervals or are in the process of moving towards more frequent reporting timeframes. For instance, Discord now publishes transparent reports on a quarterly basis, instead of semi-annually. Similarly, Twitch and Reddit started issuing transparency reports on a semi-annual basis, instead of annually.

Among the popular services that are newly included in this Report, ShareChat and Moj have the distinction of issuing monthly transparency reports. It should be noted that the new Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules (2021), abbreviated “Indian IT Rules 2021”, require from significant social media intermediaries (SSMIs) to publish monthly reports on the action taken on user complaints that they have received. As for Dailymotion, it has published only one transparency report so far but indicated its intent to do so on an annual basis (Dailymotion, 2023).

Convergence in transparency reports on TVEC stagnates, while regulatory fragmentation intensifies

Standardisation and consistency of data within a given industry are widely recognised as a best practice when it comes to transparency reporting (Radsch, 2023). The third benchmarking report showed that, while variance among the popular services’ reporting approaches remains, there had been a degree of convergence with regard to six types of reported information: proactive detection, the actions that follow a finding that there has been a TVEC violation, external reports (i.e., not by staff moderators or automated systems), TVEC-related appeals, reinstatement of content or accounts, and views of violating content.

Convergence in transparency reports on TVEC seems to be stagnating. Table 3.3 depicts convergence in the six types of information, when at least two Services are reporting metrics on one of them.

Table 3.3. Convergence in six aspects of TVEC transparency reporting

	Facebook / Instagram	YouTube	Wordpress.com	Skype / OneDrive / Teams	Twitch	Discord	Reddit	TikTok	Zoom	Snapchat	Dailymotion	Moj / ShareChat	Josh
Proactive detection	Proactive rate	Videos removed by automated flagging (number)*		Proactive rate	Proactive removal (chat messages)*	Proactive rate (servers)	Terrorist content flagged by automation (%) and removal rate	Proactive rate					
Actions following a TVEC violation finding	Content actioned (number)	Channels removed (number) Videos removed (number) Comments removed (number)	Content / sites removed as a result of an IRU notice (number and %)	TVEC actioned (number) Accounts actioned (number)	Chat messages removed (number) Channel enforcement actions (number)	Actioned reports (number and %) Accounts removed (number) Servers removed (number)	Terrorist content removals (number) Accounts suspended (number)	Videos removed (%) Accounts removed (number)*	Actioned reports (number and %)	Accounts enforced (number)	Content removed (number and %)	Takedowns (number)* Bans (number)*	Enforcement actions (%)
External reports		Human flags (number and %)	Number of IRU notices received (number)		Reports received (number)	Reports received (number and %)	User / community / external reports and/or internal escalation process (%)			Content and accounts reported (number)	Content reported (number and %)	Content and accounts reported (number and %)	

Appeals	Appealed content (number)	Removed videos appealed (number)*			Appeals (number and %)		Appeals (number)*						Grievances received (number)* Grievances pending (number)*
Content or Accounts Reinstated	Restored content (number)	Appealed videos reinstated (number)*		Reinstated accounts (%)	Appeals granted (number)	Appeals granted (number and %)		Videos restored (number)*					
Views of violating content (VVR)	Prevalence	VVR*								VVR*			

*Metric reported for all policy violations or broader categories of violating content, as opposed to TVEC-specific.

All 17 popular services that issue transparency reports covering TVEC provide information on enforcement actions. Reddit's transparency report includes a new, separate section on terrorist content removals which provides the number of pieces of content removed for promotion of terrorism and the number of accounts that were permanently suspended as a result (Reddit, 2023). Among the newly featured popular services in this report, Dailymotion provides the amount of TVEC actually removed and the percentage of the total content initially flagged or reported as TVEC that it represents, while Josh simply provides the percentage of content initially flagged or reported in the category of "terrorism and extremism" that is removed (without specifying the precise amounts). ShareChat and Moj both disclose the number of removals and bans for all categories of content. The bans can take three forms: i) *user-generated content ban*: the user is unable to post any content for a specific time period; ii) *edit profile ban*: the user is unable to edit any profile attributes for a specific duration; or iii) *comment ban*: the user is banned from commenting on any post for a specific time period.

In addition, a higher number of popular services provide information on external reports they receive from users. In the third benchmarking report, this was already the case for YouTube, Wordpress.com, Twitch, Discord, and Snapchat. Dailymotion, ShareChat and Moj now report the number and percentage of content and/or accounts reported by external sources, too. For its part, Reddit reports the percentage of potentially violative content flagged, broken down by detection method ("automation" or "other"), and the associated removal rates (Reddit, 2023).

Greater convergence regarding TVEC-specific information on appeals, as well as on content and accounts reinstated, can be observed and is strictly due to the additional metrics provided by Twitch in its latest transparency report. Along with Facebook, Instagram and YouTube, Twitch now reports the number and percentage of incoming appeals broken down by category of violating content (e.g. "Terrorism", "Violent Graphic Content", and "Hate/Harassment"). Reddit and Josh also provide information in this regard, but not specifically for TVEC. In addition, as Facebook, Instagram, Microsoft Services (Skype, Teams, OneDrive), and Discord were already doing, Twitch reports the number of appeals granted for each category of violating content. TikTok also publishes the total number of videos restored after a successful appeal, but without breaking down by type of policy violation.

The third benchmarking report highlighted that a growing number of popular services were emphasising the impact that violating content causes when it is viewed, as opposed to when it is merely posted, with more popular services reporting Prevalence and VVR metrics. In this report, no additional Service has started reporting either of these two metrics.

However, greater convergence in transparency reporting practices can be observed more generally, although it is not specific to TVEC. For instance, Twitch and TikTok have both started reporting on response times for handling user reports. In addition, following the example of Facebook, Instagram, YouTube, Microsoft Services (Skype, Teams, OneDrive), and Discord, Twitch now includes a metric on proactive detection (i.e., the number of chat messages removed proactively). TikTok provides information in this regard, as well, but it is not broken down by category of violating content.

Meanwhile, as different laws and regulations addressing TVEC and other harmful content online are coming into force across various jurisdictions, online content-sharing services increasingly face obligations to issue multiple versions of transparency reports that cover the same or overlapping reporting periods but report different metrics on the same types of content, which increases the burden and costs for services. Meanwhile, policymakers, researchers, and the public do not have access to centralised, comparable information (West, 2022).

Table 3.4 provides an overview of the regulation-specific transparency reports issued by popular services to comply with transparency obligations from several jurisdictions and which the Services happen to make available in their online transparency centres or transparency web pages (thus they are not necessarily

exhaustive). More information on some of the laws and regulations featured in the table is available in Section 5.

Table 3.4. Regulatory-specific transparency reports issued by popular services

	Austria (KoPl-G)	Germany (NetzDG)	India (Information Technology Rules)	Türkiye (Law n° 5651)	European Union (TCO)
Dailymotion				•	•
Facebook	•	•	•	•	•
Instagram	•	•	•	•	•
LinkedIn	•		•	•	•
Snapchat			•		
TikTok	•	•			
Tumblr					•
Twitch		•			•
YouTube	•	•	•	•	•
Wordpress.com					•
X		•	•	•	•

The EU Regulation 2021/784 addressing the dissemination of terrorist content online (or TCO) entered into force on 7 June 2021 and its provisions became applicable on 7 June 2022. Under this regulation, Hosting Service Providers (HSPs) have the obligation to remove terrorist content online within one hour after receiving a removal order from a competent national authority of an EU Member State; to take proactive measures when they are exposed to terrorist content; sanction platforms for non-compliance; and publish annual transparency reports (European Commission, 2023). Dailymotion, Facebook, Instagram, Twitch, Wordpress.com and Tumblr have already issued their first transparency report under the TCO. For Facebook and Instagram, Meta has implemented a specific reporting mechanism for TCO-related issues. Although X has stopped publishing transparency reports on its content moderation of TVEC, it continues to publish regulatory-specific transparency reports, as required by local laws.

Within the European Union, the Digital Services (DSA) passed into force on 16 November 2022 and its provisions comprise transparency obligations. Starting from 17 February 2024, and at least once a year, all providers of intermediary services must publish transparency reports on their content moderation practices. Further transparency requirements apply to nineteen designated very large online platforms (VLOPs) and Very Large Search Engines (VLSEs). These services must publish transparency reports at least every six months and include information on their content moderation teams, such as their qualifications and linguistic expertise (European Commission, 2023). This concerns nine of the popular services featured in this report (Facebook, Instagram, LinkedIn, Pinterest, Snapchat, TikTok, Wikipedia, X, and YouTube).

Regulatory fragmentation in transparency reporting will likely intensify as new TVEC-related laws and regulations are coming into effect. Although regulatory-specific obligations may overlap in some cases (e.g., between the DSA and the TCO), it creates an additional burden on the services that have to absorb costs and expend efforts to issue reports in multiple jurisdictions. Ultimately, the fragmentation of reported information and metrics, whether a result of companies' decisions or varying regulatory requirements, makes it more difficult to perform data comparisons and to get a cross-industry perspective, which hinders coordinated and more informed policy responses to TVEC.

Popular Services continue to rely heavily on automated means to detect and action TVEC

The first two benchmarking reports found that the popular services relied on staff member moderators, user-moderators, automated tools, or a combination of them to moderate TVEC. The third report showed a significant increase in the number of popular services using staff member moderators and automated tools. See

Table 3.5.

Table 3.5. Popular services' approaches to content moderation

Approaches	1st benchmarking report	2nd benchmarking report	3rd benchmarking report	4th benchmarking report
Services that rely on staff member moderators	40[1]	40[2]	50[3]	50
Services that rely on user-moderators	10[4]	10[5]	10[6]	9[7]
Services that rely on automated tools	At least[8] 21[9]	At least[10] 22[11]	42[12]	43[13]

As noted in the second benchmarking report, since the COVID-19 pandemic and related lockdown measures, some popular services like Facebook, Instagram, YouTube and X faced a shortage of human moderators. In response, they increased their reliance on automated monitoring systems to flag and remove problematic content, including TVEC, giving rise to concerns about human rights impacts, especially on freedom of expression and due process (OECD, 2021). The third benchmarking showed a sharp increase in the number of popular services relying on staff member moderators and automated tools due to: 1) the confirmation that all Chinese Services were using moderators and automated tools to censor content in accordance with Chinese regulatory requirements; 2) more transparency from popular services in general; 3) a higher number of members of the GIFCT and its Hash Sharing Consortium (OECD, 2022).

This report finds that popular services continue to rely on automated and hybrid approaches to detecting TVEC, and they provide additional information in this regard. For instance, Pinterest explains that it uses machine learning models to assign scores to content added to the platform. The automated tools can then use these scores to perform appropriate enforcement actions. In Q4 2022, 99% of pins deactivated for the category "Violent Actors" (in which TVEC is included) were done so by hybrid tools (Pinterest, 2023). YouTube indicates that between April and June 2022, approximately 95% of the videos removed for violating its "Violent Extremism policy" were first automatically flagged. Dailymotion provides detailed explanation on its use of automated tools such as dynamic lists of keywords and short sentences, fingerprinting technology, and hash technology. And Discord explains that in Community Servers larger than 200 members, it now takes a more proactive and automated approach to safety and may use automated means to detect violations of its policies.

In its transparency report covering H1 2022, Twitch recorded an overall decrease in user reports by 10 to 15% in all categories, due at least in part to new technologies and tooling, which are successfully deterring more reportable behaviour proactively. For "Terrorism", however, user reports increased by 144%, but Twitch explains that this appears to be largely attributed to the Russian Federation's aggression against Ukraine. Weekly reporting for "Terrorism" increased from 4.5K reports/week prior to the beginning of the conflict (24 February 2022) to 12.1K reports/week thereafter (+166%). Twitch also launched a new machine learning model to catch violative usernames and automatically enforce against them or prompt them to reset, depending on the severity of the violation.

Some of the Chinese Services disclose more information on their use of automated tools. For example, Bilibili uses automatic comparison, labelling and screening to detect content that violates its Community Guidelines. Additionally, it launched a self-developed AI system in 2021, the "Avalon Community Self-Purification System", designed to analyse users' intentions and behaviours and intercept negative content. In 2021, the Avalon System automatically processed over 720,000 pieces of negative content per day (Bilibili, 2021).

All three Indian Services also use proactive detection methods to moderate TVEC on their platforms. Josh explains that its in-house machine learning product and processes are central to the review process, and help detect and remove violating content, including TVEC. ShareChat and Moj also offer extensive explanations on their use of automated tools: every post created on their platforms is examined by a series of AI models to obtain a risk score of a post being in violation of their Community Standards, and based on the score, possibly flagged for human review. Automated content moderation tools used by ShareChat and Moj include, for example, label propagation (i.e. propagating the labels from a few sets of labelled examples to unlabelled examples based on their similarity to the labelled examples), active learning, semi-supervised learning, classifiers, and vision models to process images and videos, etc. (Mandav, Parihar, Saket, Gupta, & Mukherjee, 2021).

Various popular services have introduced or are developing Automods, which are automated tools that assist with content moderation. When the third report was written, Twitch and Reddit already had an Automod in place. In June 2022, Discord introduced its own AutoMod, a customisable moderation tool equipped with a keyword filter that can automatically detect, block, and alert users of messages containing harmful words or phrases before they are posted in text channels, threads, or Text Chat in Voice (Discord, 2022). In the same vein, Wikimedia's Moderator Tools team started working on an Automod tool for moderating content on Wikipedia automatically, based on machine learning. In contrast, Wordpress.com continues to moderate content only manually (Automattic, 2023).

In some cases, the increasing reliance on automated tools is implemented at the expanse of human moderation. For instance, in May 2022, individuals became no longer eligible to YouTube's ten-year old "Trusted Flagger" programme. Following that, in June 2022, YouTube rebranded the "Trusted Flagger" programme to the "Priority Flagger" programme, which provides policy training and tools for government agencies and NGOs to notify YouTube of content that may violate its Community Guidelines (YouTube/Google, 2023). This change comes at a time when YouTube is increasingly relying on automated tools for content moderation. YouTube has indicated that the platform is moving away from individual flaggers because it feels its AI systems can take over their workload (Hale, 2022).

While automated moderation tools can help reduce the burden of repetitive and disturbing labour on human moderators, without proper human oversight, they can also have a negative impact on human rights, notably freedom of expression. For example, Facebook and Instagram's transparency reports show a spike in content actioned for terrorism between Q1 and Q2 2023. For Facebook, this number rose from 923,000 to 1 million pieces of content; and for Instagram, from 1.6 million to 2 million pieces of content. The reports specify that this can be partly explained by an increase in enforcement of non-violating content due to a bug in Meta's proactive detection technology. As a result, over the same period, appealed content for terrorism increased by 122% for Facebook, and by 176% for Instagram. Although the bug was later fixed and the content restored, it illustrates the risks of "false positives" and undue censorship that automated moderation can bring.

As in previous reports, this also confirms the importance of human intervention to ensure fair and effective content moderation, as well as the protection of human rights and fundamental freedoms. Human judgement remains necessary to review highly contextual, nuanced content (OECD, 2022). The central role of human moderators, for example in appeal systems, underlines the need to rethink their place within

moderation processes, inasmuch as they are often unseen, poorly paid, contingent, and under pressure to perform quickly and accurately, in addition to being exposed to disturbing content (Stackpole, 2022).

Furthermore, when assessing the violative nature of potential TVEC, a growing share of popular services takes offline behaviour into account. Previously, Facebook, Instagram, Zoom, TikTok, and Twitch were already considering offline behaviour. In addition to those, Discord states that it takes off-platform harmful behaviour into consideration when assessing whether an account has violated a specific Community Guideline after becoming aware of highest-harm threats (including organising, promoting, or supporting violent extremism, making threats of violence, carrying out acts of violence, or sexualising children). Again, X is an exception: its new policies do not mention offline activities anymore, whereas previously, it clearly stated that both on- and off-platform activities were examined to determine whether a group was engaging in and/or promoting violence.

Lastly, in the first benchmarking report, obtaining a clear understanding of whether content was reviewed proactively and/or reactively was difficult with respect to 22 popular services. This number declined to 17 in 2021, and to 5 in 2022, through the combined effect of more popular services disclosing the use of automated technologies to filter and block content, and the confirmation that all Chinese platforms use such technologies. In this report, this number drops to 2.

Notification and appeal mechanisms and processes

The Santa Clara Principles on Transparency and Accountability in Content Moderation (Santa Clara University's High Tech Law Institute, 2021) offer a number of basic standards to ensure that content moderation does not unduly interfere with users' human rights and fundamental freedoms, including notifications of enforcement decisions and the possibility to appeal them. The third benchmarking report found that about three fifth of the popular services both notified their users in case of potential policy violations and had appeal processes in place.

Since then, as evidenced in Table 3.6, the number of popular services that provide notifications of enforcement decisions as well as the number of popular services that give users the right to appeal enforcement decisions increased from 30 to 38 and from 31 to 40, respectively.

Table 3.6. Popular services' approaches to notifications and appeals

Approaches	1st benchmarking report	2nd benchmarking report	3rd benchmarking report	4th benchmarking report
Services that have mechanisms for notifying users in case of potential violations of their ToS and other governing documents	21[14]	21[15]	30[16]	38[17]
Services that have appeal processes in place in respect of content moderation decisions and other measures applied under their governing documents	23[18]	24[19]	31[20]	40[21]

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

Disclosures by Chinese platforms

The previous reports explained that the Chinese Services' limited disclosures regarding content moderation and monitoring likely reflected the need to strike 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, Ruan, Crete-Nishihata, & Deibert, 2018), 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, Ruan, Crete-Nishihata, & Deibert, 2018), such as the June 2017 cybersecurity law, which among other things mandates the immediate removal of banned content and obliges Internet-based services to assist security agencies with investigations (Creemers, 2018)

On the one hand, since the last benchmarking report, some of the Chinese platforms made an effort to be more transparent about their policies and content moderation practices. WeChat provides clear definitions of "terrorism", "violent extremism", and "hate speech" in its Community Guidelines, acknowledges the use of AI tools for content moderation, and provides a list of factors considered when taking action against a violating user (i.e., severity of the breach; intentionality; legality; history of violations) (WeChat, 2023). Kwai, known as Kuaishou in China, defines "terrorist organisations" and indicates that it uses machine learning technology to review content on a large scale (including videos, comments, livestreams, and advertising). If a violation is suspected, it is then reviewed by a team of local moderators. Bilibili explains that all content, particularly content flagged by the AI screening system, is manually checked. As of the end of 2022, its content audit team totalled 3,874 employees. Bilibili also provides training for content auditors and conduct assessments through its online platform. Baidu, the parent company of Baidu Tieba and iQIYI, discloses detailed information on its content moderation systems, which leverage AI technology to preliminary filter and remove illegal content such as texts (multi-mode search and matching based on a million-word illegal entries vocabulary), images (recognition based on a million-picture prohibited image database), videos (using key frame extraction, and ASR audio-to-text conversion, voiceprint recognition, etc.). Suspected harmful content that is difficult for algorithms to judge is manually reviewed (Baidu, 2022). Besides, Xigua Video set up an appeal mechanism available to users who do not agree with a content moderation decision. On the other hand, the remaining Chinese platforms, namely Douyin, QQ, Toutiao, Youku Tudou, and QZone, disclose very limited information.

Chinese social media platforms and apps still play a paramount role in the implementation of China's "social credit system", largely deemed a mass surveillance and governmental control system in Western societies (Lix Xan Wong & Shields Dobson, 2019). Services that do not comply with local laws and regulations may face financial sanctions from the Cyberspace Administration of China. Among others, Douban, Weibo, Baidu Tieba, and QQ were fined millions of yuan for allowing "illegal" content to be seen on their platforms. In 2022, WeChat and Weibo, both owned by Tencent, banned multiple accounts from users who shared images of a pro-democracy demonstration (Yang, 2022). That same year, Weibo announced that it would start publishing users' province or municipality (for users in China) and IP locations (for users overseas) on their account pages and when they post comments, with a view to "reduce bad behaviour" (Reuters, 2022). As for Xiaohongshu, a leaked document revealed how the platform deals with censoring discourse about "sudden incidents". The document lists a wide range of particularly sensitive content that requires special treatment, including terrorist attacks, mass incidents, or fatal stabbings, but also demonstrations, student strikes, or widespread public criticism of government institutions, that must

be reported to community managers and the Shanghai Operation Security Group within five minutes (Boyd, How Xiaohongshu censors "sudden incidents", 2022).

Particular scrutiny is applied to livestreamed content. In 2022, the National Radio and Television Administration and Ministry of Culture and Tourism jointly published guidance for the live-streaming industry, listing 31 banned behaviours and setting expected standards of quality and content for both livestream hosts and platforms. Banned behaviours range from publishing content promoting terror and exaggerate violence and spreading false terrorist information, to posting content that denies the leadership of the Communist Party of China or “denigrates the fine cultural traditions of the nation” (Tindall, What are the current regulations for live streaming in China?, 2022). ByteDance, the parent company of TikTok, Douyin, Xigua Video, and Toutiao, automatically transcribes livestreamed videos and audio clips into text and uses algorithms to detect any sensitive terms, dates, or names (Lu, I helped build ByteDance's censorship machine, 2021).

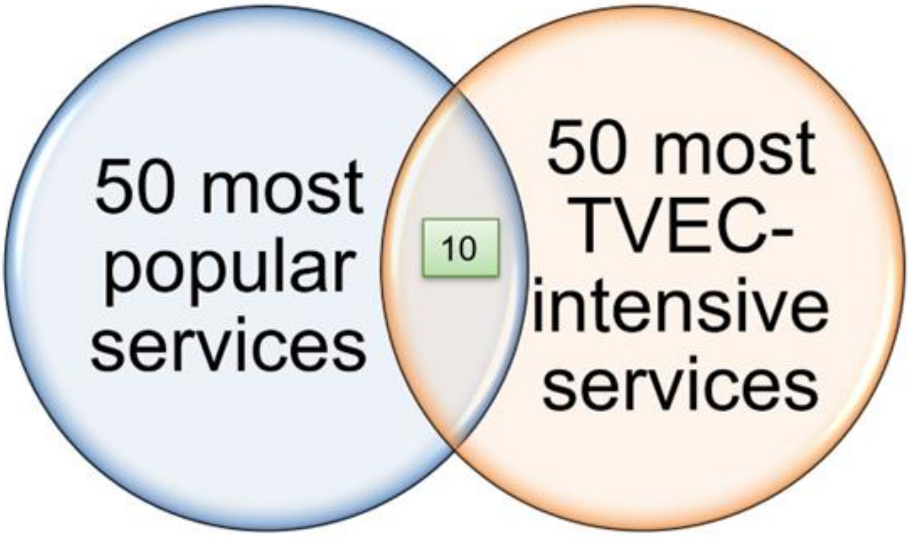
In this context of strict surveillance, content moderators in China are overworked and chronically stressed. For example, testimonies from former employees of Bilibili reveal that their job required them to monitor as many as 1,600 video clips in a 12-hour shift and to work overtime to the detriment of their physical and mental health (Meihan, China's content moderators are overworked and chronically stressed, 2022).

Finally, to escape censorship, an increasing number of Chinese users turn to other platforms, such as Reddit, drawn by the community-moderated, discussion-based format which allows more fringe voices to thrive. As Reddit is banned in China, users in the country have to use virtual private networks (VPNs) to thwart the Great Firewall (Chen, 2023).

4 Updated commonalities, developments and trends in the intensive services' approaches to TVEC

Turning to the top 50 Services that are most widely used for accessing and spreading TVEC online, as in the third benchmarking report, the composition of the group is markedly different from the top 50 most popular services. Figure 4.1 shows that only ten services appear on both lists. They are Discord, Dropbox, Facebook, Google Drive, Instagram, Telegram, TikTok, WhatsApp, X, and YouTube. In the previous edition, 11 services were on both lists: the same 10 Services, with the addition of VK, which is not featured in the present edition.

Figure 4.1. Overlap between the popular and intensive services lists



Despite small progress made by some intensive services to improve their governing documents, the majority still does not explicitly prohibit TVEC

Similar to what was observed in the third benchmarking report, Table 4.1 shows that only 40% of intensive services specifically prohibit TVEC in their governing documents. This reflects a slight change, from 18 in 2022 to 20 in 2024.

Table 4.1. TVEC definitions in intensive services' governing documents

Degree of elaboration of TVEC definitions	3rd benchmarking report[22]	4th benchmarking report
Governing documents that explicitly ban the use of their technologies to foster terrorist and/or violent extremist aims and define terrorism, violent extremism and related concepts with sufficient detail to understand the scope of such terms, providing examples where appropriate	6[23]	12[24]
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 extremist/violent extremism and similar expressions	12[25]	8[26]
Governing documents that use broad and/or vague descriptions of prohibited conduct, which descriptions can be interpreted as supersets encompassing TVEC	18[27]	14[28]
No content prohibition / Terms of Service	13[29]	16[30]

Note: The figures in the column of the 3rd benchmarking report are an aggregation of the three categories used in the previous edition: i) mainstream TVEC-intensive Services; ii) file-sharing TVEC-intensive Services; and iii) far-right-focused TVEC-intensive Services. The detailed list of Services for each category is provided in the endnotes.

Among the intensive services that expressly prohibit TVEC in their governing documents, slight advances can be noted towards more precision and clarity in the definitions and explanations provided. In the third benchmarking report, 12 intensive services explicitly banned the use of their technologies to foster terrorist and/or violent extremist aims, using (but not explaining in detail) the terms terrorist/terrorism, violent extremist/violent extremism, and similar expressions. In the present report, this group decreased in size from 12 to eight, while the group that bans and defines TVEC with sufficient detail to understand the scope of the terms doubled in size from six to 12.

However, that increase is largely due to the inclusion of Services in this edition's Intensive list that were not on the third report's Intensive list. Therefore, the apparent progress is overstated because it is mainly the result of newly featured intensive services in this report (namely Rocket.Chat, Slack, SoundCloud, Threads) already prohibiting TVEC and explaining the key terms clearly. For example, SoundCloud give specific examples of content that is allowed or not under its "terrorist content" policy (SoundCloud, 2023).

As among the popular services (see Section 3), an increasing share of intensive services rely on designation lists to define TVEC. This number grew from two in the third benchmarking report to four in the present report. For instance, SoundCloud explains that, while it takes into consideration the United Nations', the United States' and the European Union's lists of terrorist designated organisations, it also follows and enforces the details Pursuant to Article 3 of Regulation (EU) 2021/784 of the European Parliament and of the Council (SoundCloud, 2023). In addition, both Slack and Rocket.Chat prohibit the use of their services to "provide material support or resources (or to conceal or disguise the nature, location, source, or ownership of material support or resources) to any organisation(s) designated by the United States government as a foreign terrorist organisation pursuant to section 219 of the Immigration

and Nationality Act or other laws and regulations concerning national security, defence or terrorism" (Slack, 2021) (Rocket.Chat, 2023). Alongside these newly featured intensive services, as in the previous report, Rumble continues to prohibit content or material that promotes or supports entities and/or persons designated by either the Canadian or United States government as terrorists or terrorist organisations (Rumble, 2023).

Despite this qualified progress, 60% of the intensive services still do not even explicitly prohibit TVEC, let alone define it clearly. More specifically, 14 intensive services use overly broad and/or vague descriptions of prohibited conduct, which descriptions could be interpreted as supersets encompassing TVEC (against 18 in 2022), and 16 Intensive Services provide no information at all (against 13 in 2022).

Half of the intensive services remain opaque regarding their approaches to content moderation

Furthermore, as shown in Table 4.2, and as was also the case in the third benchmarking report, some intensive services explain their approach to content moderation in detail, while others do not communicate any information in this regard.

Table 4.2. Intensive services' content moderation approaches

Degree of elaboration of content moderation approaches	3rd benchmarking report[31]	4th benchmarking report
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)	13[32]	18[33]
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")	12[34]	7[35]
Vague statements on content moderation (e.g. "we have the right but not the obligation to remove content…")	3[36]	8[37]
Provision of contact form only	4[38]	1[39]
No information on content moderation available	17[40]	16[41]

Note: The figures in the column of the 3rd benchmarking report are an aggregation of the three categories used in the previous edition: i) mainstream TVEC-intensive Services; ii) file-sharing TVEC-intensive Services; and iii) far-right-focused TVEC-intensive Services. The detailed list of Services for each category is provided in the endnotes.

A minority of Intensive Services provide detailed information on their content moderation practices. For instance, Element now provides a more precise list of the different actions it may take in case of breaches of its Terms of Use (e.g., temporary or permanent withdrawal of the user's right to use the platform, temporary or permanent removal of content, warning, legal action, disclosure of information to law enforcement authorities) (Element, 2023).

Among the intensive services that are included in this report for the first time, some of them provide detailed explanations about their approaches to content moderation. For example, Mastodon explains that, in 2022, it added the ability to quickly suspend all accounts matching specific search queries, such as a matching IP range or email domain. Mastodon also introduced a webhook system to allow server operators to setup more elaborate automation systems for their moderation needs and expanded the set of moderation APIs to support it (Mastodon, 2023). Moreover, SoundCloud states that it may use automated and manual tools

to detect and prevent objectionable content, such as violative profile names, URLs, and profile descriptions, but also violative audio, images, and track data. Lastly, Services that appear on both the popular services list and the intensive services list, such as Discord, Dropbox, Facebook, Instagram, Google Drive, TikTok, or YouTube, describe the different types of sanctions that may be taken in case of TVEC being identified on their platforms, as well as Services who are just on the intensive services list, such as SoundCloud or Threads.

In the third benchmarking report, 24 out of 50 intensive services either provided vague statements on their content moderation approaches, only provided a contact form, or did not provide any information at all. This number remained quite stable, with 25 intensive services currently falling into either one of those three categories. Overall, half of the intensive services remain non-transparent regarding their approaches to content moderation: eight provide vague statements, one only provides a contact form, and 16 provide no information at all. For these latter, it is difficult to assess which ones have content moderation processes in place.

The majority of intensive services either have no notifications and appeal mechanisms in place or do not provide any information

Since the third benchmarking report, virtually no progress has been made regarding the implementation of notifications and appeal mechanisms. On the contrary, the figures in Table 4.3 indicate a decline in the number of intensive services with such processes in place. Currently, 11 intensive services notify users in case of potential violations, against 16 in 2022. Likewise, 16 intensive services have made appeal processes available to users to dispute content moderation decisions, against 18 in 2022. However, this decrease is not due to services withdrawing existing notifications and appeal mechanisms, but rather to changes in the composition of the intensive services list.

Table 4.3. Intensive services' notifications and appeals mechanisms

Notifications / appeals in place	3rd benchmarking report	4th benchmarking report
Intensive Services that have mechanisms for notifying users in case of potential violations of their ToS and other governing documents	16[42]	11[43]
Intensive Services that have appeal processes in place in respect of content moderation decisions and other measures applied under their governing documents	18[44]	16[45]
No notifications and appeals specified / no information available	30[46]	32[47]

Note: The figures in the column of the 3rd benchmarking report are an aggregation of the three categories used in the previous edition: i) mainstream TVEC-intensive Services; ii) file-sharing TVEC-intensive Services; and iii) far-right-focused TVEC-intensive Services. The detailed list of Services for each category is provided in the endnotes.

In the absence of any governing documents or publicly available information, it is impossible to determine whether the Services have notifications and appeal mechanisms in place. This raises questions about these Services' ability to ensure due process, and to guarantee the respect of human rights and fundamental freedoms.

Transparency reporting on TVEC among intensive services decreases

As shown in Table 4.4, only six of the 50 intensive services identified issue transparency reports that specifically cover TVEC, against 8 in the third benchmarking report. The vast majority (five out of six) are mainstream platforms and are also included in the popular services list (see Annex A). For this reason, the content of their transparency reports is discussed in Section 2 above.

As noted in Section 3 of this Report, X (previously Twitter) has stopped publishing transparency reports and the last one available covers July to December 2021. Mega is the other service that was among the eight intensive services that were found to specifically report on TVEC in the third edition but is not included in the same category in this report. The reason, however, is not that it no longer reports on TVEC, but that it is no longer considered to be an Intensive Service. As a result, given that no Service that is new to the Intensive list has started publishing transparency reports specifically covering TVEC, the overall number of intensive services that do so has decreased.

Table 4.4. Intensive services that issue transparency reports with TVEC-specific information

3rd benchmarking report	4th benchmarking report
8	6
Discord	Discord
Facebook	Facebook
Instagram	Instagram
Justpaste.it	Justpaste.it
TikTok	TikTok
YouTube	YouTube
Twitter	
Mega.nz	

Outside of the mainstream platforms that are already included in the popular services list, Justpaste.it is the only Intensive Service to issue transparency reports expressly addressing TVEC.

Justpaste.it is a text- and images-sharing platform that caters to privacy-sensitive users. As the platform is anonymous by default, users do not have to create an account to publish on the website, so it is unnecessary to disclose any personal information (name, email, or home address) (Justepaste.it, 2018). The Islamic State took advantage of such features to disseminate TVEC on Justpaste.it. Subsequently, Justpaste.it collaborated with law enforcement to remove the TVEC uploaded on the platform.

Since 2019, to strengthen user trust and safety on its platform, Justpaste.it has also been publishing annual transparency reports that contain information on TVEC. In its latest transparency report (covering 2022), Justpaste.it reported the number of requests received from governments and law enforcement agencies regarding illegal content (broken down by EU requests, UK requests, and the Republic of Türkiye (hereafter "Türkiye" 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, 2022).

Justpaste.it's commitments to combat TVEC and improve transparency is to be commended, given the company's small size. Its founder, Mariusz Żurawek, is its only employee (Ilinsky, 2019). Hence,

Justpaste.it continues to prove that it is not infeasible for small Services to establish and implement content moderation policies and procedures or to publish regular transparency reports (OECD, 2022).

As explained in the third benchmarking report, two main factors may explain the differences among the intensive services, whether in the degree of elaboration of their governing documents, or in their efforts to detect and remove TVEC, ensure due process, and publish transparency reports. First, these Services have substantially different financial resources at their disposal. Ten of them are also included in the top 50 popular services list (see Annex A) and generate consistent and significant profits from a wide range of activities. These large intensive services tend to have more robust governing documents and be better equipped to identify and action TVEC, whether through automated tools or dedicated staff (OECD, 2022). At the other end of the spectrum, small and micro platforms may lack the financial and human resources to establish appropriate Terms of Service and Community Guidelines, deploy content moderation tools, hire human moderators and Trust & Safety teams, establish notifications and appeal processes, publish transparency reports, or conduct risk assessments. As a result, as explained in the Introduction, terrorists and violent extremists proliferate on less capable, less moderated, fringe platforms (OECD, 2022).

Second, some of the intensive services have little or no interest in moderating content and preventing the dissemination of TVEC on their platforms. Some are self-proclaimed defenders of freedom of speech and pride themselves on providing spaces where all views can be expressed, even the most hateful ones (OECD, 2022). Typically, these Services identify themselves as free speech advocates and alternatives to "Big Tech" giants and their perceived over-censorship (Stocking, et al., 2022). For instance, Gab states in its Terms of Service that it strives to ensure that the First Amendment remains the website's standard for content moderation (Gab, 2023). In a similar vein, Rocket.Chat explains that it wants to be "a platform that allows for free and unrestricted communication" (Rocket.Chat, 2020). For its part, Rumble writes that amid "the recent rise of cancel culture", it supports "diverse opinions, authentic expression, and the need for open dialogue" (Rumble, 2024). 4chan has styled itself for several years as a "bastion of free speech", where "politically incorrect" and extreme views can thrive, guaranteed by the complete anonymity of users, as well as the fast-changing and ephemeral nature of the content it hosts (Provetti, 2021). Likewise, the video streaming platform Odysee has become a safe haven for violent far-right content creators and constitutes an attractive option for users that have been banned or demonetised from larger platforms (Leidig, 2021). When asked about videos on Odysee promoting the Proud Boys, a militant far-right group whose leaders were convicted of seditious conspiracy for their role in the 6 January 2021 attack on the Capitol in Washington, D.C., Odysee's CEO responded that such groups "should be allowed to speak to others that want to hear them" (Fernandez-Aubert, Reinhart, & Squire, 2023).

Finally, many intensive services are operated by terrorist and violent extremist actors or sympathisers. Such platforms and websites are specifically created to generate and spread TVEC. This is the case, for example, of "jihadi"-operated websites like Shahadanews.com (al-Shabab's news website), Amjaad.video (a video platform run by Hayat Tahrir al-Cham) or Dalelansar.info (a website controlled by the Islamic State). Other websites like LiveGore.com, Americanfuturistpublishing.com, and 3pdirectory.com actively support neo-Nazi rhetoric, white supremacy, and other violent extremist ideologies.

5 Enacted and emerging TVEC-related laws and regulations

Governments initially relied mainly on self-regulation and voluntary pledges from Internet-based services for addressing TVEC online. However, industry efforts to counter TVEC have been perceived as inadequate, thus triggering a shift towards increased government oversight (Gorwa, 2019). In particular, a growing number of governments have proposed and enacted laws and regulations to thwart the spread of TVEC online. This Section provides an overview of those laws and regulations in several OECD jurisdictions.

Australia

The Online Safety Act 2021 reforms and expands existing online safety regulations in Australia, consolidating and modernising separate regulatory regimes. It updated the schemes on cyberbullying, image-based abuse, and illegal and restricted online content, and introduced 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 (if these codes do not meet community expectations, eSafety can issue industry standards); and
- Require online service providers to report on how they are meeting the Basic Online Safety Expectations, which includes 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 are accessible to end-users in Australia and 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 a blocking 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.

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, 2021) 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

The industry associations indicated they would adopt the two-phase approach to codes development outlined in eSafety's position paper. The first phase focused on class 1 material, including child sexual exploitation material and terrorist material. eSafety considered the draft industry codes submitted by the industry associations and, on 9 February 2023, the eSafety Commissioner advised that the class 1 draft industry codes did not provide appropriate community safeguards and were unlikely to meet the statutory requirements for registration. The eSafety Commissioner invited the industry associations to respond and/or resubmit draft industry codes that addressed eSafety's feedback.

On 16 June 2023, the eSafety Commissioner registered the Social Media Services Code, Apps Distribution Services Code, Hosting Services Code, Internet Carriage Services Code, and the Equipment Code. After addressing concerns from the eSafety Commissioner regarding generative AI and its integration with search engines, the Search Engines Services Code was registered on 12 September 2023. The eSafety Commissioner made the decision not to register Relevant Electronic Services Code and the Designated Internet Services Code drafted by the online industry as they failed to provide appropriate community safeguards to deal with illegal and harmful content online. eSafety will now move to develop mandatory and enforceable industry standards for Relevant Electronic Services and Designated Internet Services.

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 two rounds of non-periodic reporting notices requiring information from specific providers about how they are meeting certain expectations relating to child sexual exploitation and abuse. eSafety has published transparency reports containing information received in response to these notices. These transparency reports are available on the eSafety website.

eSafety has also issued a non-periodic reporting notice requiring information about how a specific service provider is meeting certain expectations relating to online hate. This regulatory process is still underway as of December 2023 and eSafety intends to publish appropriate information in due course.

Austria

The Communication Platforms Act (*Kommunikationsplattformen-Gesetz* or “KoPI-G”) (Government of Austria, 2021) is a federal act providing measures to protect users on communication platforms and is part of a larger regulatory package against online hate speech. The KoPI-G came into force on 1 January 2021.

The Act applies to domestic and foreign providers of for-profit communication platforms that have more than 100,000 users or an annual revenue exceeding 500,000 euros, with exemptions for online encyclopaedias (such as Wikipedia) and learning platforms, newspaper and television company platforms hosting their journalistic offerings, and apps used for individual communication.

The KoPI-G imposes the following obligations on platforms:

- Implement an "effective and transparent procedure" for reporting and deleting illegal content.
- Take down certain types of illegal content within 24 hours (the law defines this as content whose “illegality is already evident to a legal layperson”) and otherwise unlawful content within seven days after a complaint has been filed. Takedowns, including the content itself, as well as information identifying the user and date/time, must be archived for 10 weeks to retain evidence for possible prosecution.
- Introduce comprehensive due process reporting systems for contesting takedowns, but also for users to report content and to inquire in case something they reported has not been deleted. Appeals are made to the platform, but complaints can be issued to the RTR Austrian Regulatory Authority for Broadcasting, which has as its independent supervisory body the Austrian Communications Authority (KommAustria).
- Publish reports on takedowns, on an annual basis if they have more than 100,000 users, and on a quarterly basis for those exceeding 1 million users. Platforms have to provide a description of “the organisation, personnel and technical equipment, technical competence of the staff responsible for processing reports and review procedures, as well as the education, training and supervision of the persons responsible for processing reports and reviews”.
- Appoint and notify the authority of (a) an authorised representative who is responsible for compliance with the new laws, and (b) a representative who can accept official communications on behalf of the platform.

- Pay a financing contribution to the budget of the regulatory authority.

If platforms fail to comply, they can be subject to fines of up to 10 million euros (depending on several factors such as revenue, number of users, prior misconduct, severity and length of violation).

Canada

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

Mandate letters sent by Canada's Prime Minister to the Minister of Justice and the Minister of Canadian Heritage on 16 December 2021 designated the introduction of legislation 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, Minister of Canadian Heritage Mandate Letter, 2021) (Trudeau, Minister of Justice and Attorney General of Canada Mandate Letter, 2021).

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[48].

European Union

The "Regulation to address the dissemination of terrorist content online" was adopted on 28 April 2021 and entered into force on 7 June 2022. 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 Directive (EU) 2017/541 on combating terrorism (European Parliament; European Council, 2017), 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. 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 Digital Services Act (DSA) was adopted on 19 October 2022. The DSA contains provisions applicable to all providers of online platforms, as well as additional ones for "very large online platforms" (VLOPs, with more than 45 million users per month in the EU). The DSA imposes new obligations on digital service providers centred around four main principles:

- ***Transparency***: All digital service providers of intermediary services (which means both providers of hosting services and providers of online platforms) must publish clear, comprehensible and detailed annual reports on content moderation (with additional information required for online platforms and VLOPs. Online platforms must provide transparency on advertisements (e.g. clearly signal advertisements; provide information on the natural or legal person on whose behalf the advertisement is presented or who paid for it; and provide information on parameters used to determine the recipient) and on the algorithms used to display them (with additional requirements for VLOPs). VLOPs must also publish information on their use of recommender systems. Online platforms, such as online marketplaces, must also ensure that traders provide sufficient information to the platform and display trader information to users.
- ***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. They must also inform the relevant authorities if they suspect a serious criminal offence involving a threat to the lives or safety of persons. VLOPs must also take steps to manage systemic risks, including annual risk assessments, risk mitigation measures, annual independent audits and the appointment of compliance officers.
- ***Industry co-operation***: The European Commission will support and promote the development of voluntary industry standards, codes of conduct and crisis protocols for 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 concerning:

- 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.

Starting from 17 February 2024, all providers of intermediary services must publish transparency reports on their content moderation practices.

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 Germany's NetzDG law. 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 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 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

- clarifying whether the provider has given access to their data to independent researchers
- service providers that receive more than 100 reports about unlawful content per calendar year must publish bi-annual reports on their reports handling process.

Non-compliance with 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, 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 bi-annual 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

In Ireland, the Online Safety and Media Regulation Act 2022 applies to all "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 comes 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 Act is the first piece of Irish legislation that deals with the regulation of video-sharing platforms, including YouTube. The Act 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 Act establishes a Media Commission, including an Online Safety Commissioner. This new body replaces the Broadcasting Authority of Ireland and is also responsible for the regulation of on-demand services, including radio, television, and video-on-demand services.

Under the Act, online services must 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
- submitting reports regarding their compliance with the Online Safety Codes.

The Act sets out four categories of harmful online content:

- material that is a criminal offence to disseminate under Irish or EU law (for example, child sexual abuse material or terrorist content);
- cyberbullying material;
- material encouraging or promoting eating disorders; and
- material encouraging or promoting self-harm or suicide.

The Media Commission can also include or exclude other categories of harm.

The Media Commission has 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 has 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 Parliament of New Zealand enacted the Films, Videos and Publications Classification (Urgent Interim Classification of Publications and Prevention of Online Harm) Amendment Act 2021 (Government of New Zealand, 2021).

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

Under the Act, 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 Act also authorises an Inspector of Publications to issue a takedown notice for objectionable online content. Such notices are issued to an online content host and direct the removal of a specific link to make it no longer viewable in New Zealand. Failure to comply can result in civil pecuniary penalties.

Furthermore, the Act 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[49] (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 Act 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 established a regulatory unit to respond to reports of TVEC online, which relies on voluntary co-operation to remove TVEC. The Act also enables future mechanisms for blocking or filtering TVEC that is deemed to be objectionable in New Zealand, should this become necessary. The Act requires that a very clear governance and reporting system underpin any such filter.

Korea

Korea has passed several anti-terrorism laws that cover online material. Korean legislation allows the head of a related agency (i.e. a State agency engaged in counter-terrorism activities, a local government, and any other agency prescribed by Presidential Decree) to request the co-operation 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 co-operation 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 (Tech Against Terrorism, 2021). 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, the obligations of social network providers have been first introduced in 2020 by additional Article 4 of Law No. 5651 on Regulation of Publications on the Internet and Combatting Crimes Committed by Means of Such Publications Law. Subsequently, the scope of this Article was further expanded in 2022. With the recent amendments, the responsibilities of social network providers in Türkiye have been redefined as delineated below.

The Article mandates foreign-based social network providers with over one million daily accesses from Türkiye to appoint a representative. This is a strategic move to ensure compliance with Turkish law, facilitating direct communication with Turkish authorities. In cases where daily accesses exceed ten million, the appointed representative shall be authorised adequality in technical, administrative, legal, and financial matters by the social network provider, thereby ensuring a more accountable presence to administrative and judicial authorities as well as users in Türkiye.

A significant aspect of the amendments is the introduction of a tiered penalty system for non-compliance. Initially, failure to appoint a representative lead to administrative fine. If the non-compliance persists, the penalties escalate, ranging from additional fines to advertising ban, and in extreme cases, bandwidth throttling by up to 90%. This progressive approach aims to ensure compliance while providing a pathway for rectification.

The amendments place a strong emphasis on prompt responsiveness to content-related complaints. Social network providers are required to address complaints regarding content that infringes personal and privacy rights within 48 hours. This rapid response mechanism aims to safeguard user rights and uphold the integrity of online discourse, ensuring that it does so without infringing upon the fundamental rights of users in Türkiye.

Further, the amendments introduce reporting obligations in order to increase transparency and maintain accountability of platforms. Social network providers must submit bi-annual reports in Turkish, covering compliance with their legal obligations. These reports have become comprehensive, now including information on content visibility algorithms, advertising policies, and measures against unlawful content taken by them. This objective to increase transparency and maintain accountability is a key component in the regulation, providing legal safeguards to users in Türkiye.

In response to serious crimes such as child sexual abuse, social network providers must cooperate with judicial authorities by sharing information that could help identify offenders. This collaboration is crucial for ensuring public safety and fighting against efforts of mal-intended entities to disseminate illegal content online.

Overall, these amendments reflect Türkiye's proactive stance in regulating social network providers. By imposing these requirements, Türkiye aims to safeguard personal rights, enhance public safety, and ensure that social network providers operate transparently and responsibly within its jurisdiction.

In Türkiye, Law No. 6112 regulates vide radio and television broadcasting services and o on-demand media services transmitted by any and all techniques, methods or means and by electromagnetic waves or other means under any denotation. Under this law, media service providers must get a licence from Turkish authorities (namely, the Radio and Television Supreme Council). Turkish authorities have the power to remove inappropriate content, which violates the above-mentioned Law and related by-Laws, 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). The plan was 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.

The Online Safety Act 2023 applies to providers of i) online platforms that allow user to generate, upload, or share content with others, and ii) search services. Online platforms or search services fall within the scope of the legislation if they: i) have a significant number of users in the UK, ii) target the UK market, or iii) otherwise present a material risk of significant harm to individuals in the country.

The Act requires those services falling within its scope to (Government of the United Kingdom, 2023):

- 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 can be either physical or psychological. Illegal content includes content that amounts 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, affected companies will be bound, among other things, to:

- 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 TVEC. 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 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.

6 Conclusion

In view of the foregoing, there is no single solution to stop the dissemination of TVEC online. As explained in the Introduction of this report, terrorists and violent extremists continually adapt to the transformations of the online environment, exploit new platforms and features as they emerge, and develop creative ways to escape content moderation. As a result, the TVEC landscape is multi-faceted and rapidly evolving, spreading across diverse content-sharing services, ideological lines, and geographical areas.

The paucity of transparency reporting on TVEC among online content-sharing services, Popular and Intensive alike, shows that there is much room for improvement. The results of this report also highlight the need for more precision in the Services' governing documents, when they have any; more consistency in the metrics and methodologies used to prepare transparency reports; more transparency in their approaches to content moderation; and more efforts to ensure due process and to safeguard fundamental rights and freedoms, such as the freedom of expression and the right to privacy. Transparency reporting on TVEC is paramount to better inform policymaking, assess the effectiveness of counter-measures and their potential impact on human rights. Addressing the TVEC issue in a holistic manner, supported by reliable evidence, can help reduce the volume and reach of TVEC online, prevent new terrorist attacks from being committed, and eventually save lives.

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 (Meta Platforms, Inc.)	2,963*(a)* (as of April 2023) (Datareportal, 2023)	Social networking platform	Y	Y
2	YouTube (Alphabet, Inc.)	2,527*(a)* (as of April 2023) (Datareportal, 2023)	Video streaming platform	Y	Y
3	Instagram (Meta Platforms, Inc.)	2,000*(a)* (as of April 2023) (Datareportal, 2023)	Social networking platform	Y	Y
4	WhatsApp (Meta Platforms, Inc.)	2,000*(a)* (as of April 2023) (Datareportal, 2023)	Messaging app	N	Y
5	Weixin/WeChat (Tencent Holdings Ltd.)	1,313*(a)* (as of April 2023) (Datareportal, 2023)	Social networking/content sharing/messaging platform	N	N
6	iMessage/Facetime (Apple, Inc.)	1,000*(a)* (as of February 2023) (SignHouse, 2023)	Messaging and video chat apps	N	N
7	TikTok (ByteDance Technology, Co.)	1,092*(a)* (as of April 2023) (Datareportal, 2023)	Short video app	Y	Y
8	Facebook Messenger (Meta Platforms, Inc.)	1,036*(a)* (as of April 2023) (Datareportal, 2023)	Messaging app	Y (included in Facebook)	Y

9	Zoom (Zoom Video Communications, Inc.)	810,2*(a)* (as of April 2023) (Statista, 2023)	Video chat and voice calls app	Y	N
10	Snapchat (Snap, Inc.)	750*(a)* (as of April 2023) (Datareportal, 2023)	Social networking platform	Y	N
11	Douyin (ByteDance Technology, Co.)	743*(a)* (as of April 2023) (Datareportal, 2023)	Short video app	N	N
12	Telegram (Telegram Messenger LLP)	700*(a)* (as of April 2023) (Datareportal, 2023)	Messaging app	N	N
13	Kuaishou/Kwai (Beijing Kuaishou Technology Co., Ltd)	654*(a)* (as of May 2023) (Kuaishou, 2023)	Short video app	N	N
14	QZone (Tencent Holdings Ltd.)	600*(a)* (as of December 2021) (Galov, 2023)	Social networking platform	N	N
15	Weibo (Sina Corp.)	586*(a)* (as of April 2023) (Datareportal, 2023)	Social networking platform	N	N
16	QQ (Tencent Holdings Ltd.)	572*(a)* (as of April 2023) (Datareportal, 2023)	Instant messaging and web portal site	N	N
17	iQIYI (Baidu, Inc.)	530*(a)* (as of March 2022) (Statista, 2023)	Video streaming platform (user-generated and syndicated content)	N	N
18	Pinterest (Pinterest, Inc.)	450*(a)* (as of April 2023) (Datareportal, 2023)	Social networking platform	N	N
19	Reddit (Reddit, Inc.)	430*(a)* (as of January 2023) (Curry, 2023)	Social news aggregation, web content ranking and discussion website	Y	Y
20	Dailymotion (Vivendi)	400*(a)* (as of May 2023) (Ragot, 2023)	Video streaming platform	Y	Y
21	X (X Corp. Limited)	373*(a)* (as of April 2023) (Datareportal, 2023)	Short messages-focused social networking platform	N	Y
22	Bilibili (Alibaba Group Holding Ltd.)	326*(a)* (as of Q4 2022) (Statista, 2022)	Video-sharing platform	N	N

23	LinkedIn (Microsoft, Inc.)	310*(a)* (as of February 2023) (Stern, 2023)	Jobs-focused social networking platform	N	Y
24	Baidu Tieba (Baidu, Inc.)	300*(a)* (as of September 2023) (Alexander S. , 2023)	Online communications platform	N	N
25	Douban (Information Technology Company, Inc.)	300*(a)* (as of July 2021) (Marketing to China, 2021)	Social networking platform	N	N
26	Moj (Mohallah Tech Private Limited)	300*(a)* (as of February 2023) (Roy & Mishra, 2023)	Short video app	Y	Y
27	Quora (Quora, Inc.)	300*(b)* (as of January 2023) (Bleu, 2023)	Question-and-answer website	N	N
28	Skype (Microsoft, Inc.)	300*(a)* (as of March 2023) (Wise, 2023)	Video chat and voice calls app	Y	Y
29	Toutiao (ByteDance Technology, Co.)	300*(a)* (as of September 2023)	News aggregator and content creation platform	N	N
30	IMO (PageBites, Inc.)	200*(a)* (as of November 2022) (Lyons, 2022)	Video chat and voice calls app	N	N
31	Xiaohongshu (Xingyin Information Technology Shanghai Ltd.)	200*(a)* (as of March 2022) (Zhou, 2023)	Social media platform	N	N
32	Teams (Microsoft, Inc.)	280*(a)* (as of Q2 2023) (Microsoft, 2023)	Online collaboration platform	Y	Y
33	Viber (Rakuten, Inc.)	260*(a)* (as of March 2022) (Maslar, 2022)	Messaging app	N	Y
34	Youku Tudou (Alibaba Group Holding Limited)	247*(a)* (March 2022) (Statista, 2022)	Video streaming platform (user-generated and syndicated content)	N	N
35	Twitch (Amazon.com, Inc.)	180*(a)* (as of March 2023) (Zwieglinska, 2023)	Live-streaming platform	Y	N
36	ShareChat (Mohallah Tech Private Limited)	180*(a)* (as of March 2023) (Roy & Mishra, 2023)	Social networking platform	Y	Y

37	LINE (Line Corporation)	178*(a)* (as of June 2022) (Statista, 2022)	Messaging app	N	N
38	Xigua Video (ByteDance Technology Co.)	175*(a)* (as of December 2022) (Statista, 2023)	Short video streaming app	N	N
39	Vimeo (Vimeo, Inc.)	170*(a)* (as of May 2023) (YouTube vs. Vimeo: The key differences in 2023 (updated), 2023)	Video streaming app	N	N
40	Discord (Discord, Inc.)	154*(a)* (as of January 2023)	Chat platform	Y	N
41	Josh (VerSe Innovations Private Limited)	153*(a)* (as of August 2022)	Short video app	Y	N
42	Likee (BIGO Technology Private Limited)	150*(a)* (as of December 2022)	Streaming platform	N	Y
43	PicsArt (PicsArt, Inc.)	150*(a)* (as of May 2023)	Photo and video app	N	Y
44	Tumblr (Automattic, Inc.)	135*(a)* (as of July 2023)	Microblogging and social networking platform	N	N
45	Steam	132 *(a)* (as of 2021)	Gaming platform	N	N

1. 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 is 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.)	30.28% (as of July 2023) (Datanyze, 2021)	Cloud-based file sharing	N	Y
Dropbox (Dropbox, Inc.)	21.83% (as of July 2023) (Datanyze, 2021)	Cloud-based file sharing	N	N
Microsoft OneDrive (Microsoft, Inc.)	13.96% (as of July 2023) (Datanyze, 2021)	Cloud-based file sharing	Y	Y

Name of service (parent company)	Indicative Global Market Share *(a)* or monthly average unique devices (millions) *(b)*	Type of market/service	Transparency report on terrorist/violent extremist content	Provided feedback / comments on its profile
Wordpress.com (Automattic, Inc.)	62%***(a)*** (as of July 2023) (Envisage Digital, 2021)	Content management system	Y	N
Wikipedia (Wikimedia Foundation)	2,000***(b)*** (as of June 2023) (Wikimedia, 2021)	Online encyclopaedia	N	N

Annex B. Profiles of the Top 50 Services

1. Facebook[50]

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 provides one the most comprehensive explanation of what constitutes terrorist entities and terrorist content under their Community Standards. For each category of prohibited content, Facebook explains the policy rationale and differentiates between, on the one hand, content that is not allowed and, on the other hand, content that requires additional context, that is allowed with a warning screen, or that can only be viewed by adults aged 18 and older. Facebook's Community Standards, in the section on 'Violence and Criminal Behaviour', under the 'Dangerous Individuals and Organisations' policy (Facebook, 2023), do not allow organisations or individuals that proclaim a violent mission or are engaged in violence to have a presence on Facebook. Facebook assesses these entities based on their behaviour both online and offline – most significantly, their ties to violence. Under this policy, Facebook designates 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 serial 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. Additionally, Facebook does not allow content that praises, substantively supports, or represents events that Facebook designates as terrorist attacks, hate events, multiple-victim violence or attempted multiple-victim violence, serial murders, hate crimes or violating violent events. Moreover, Facebook does not allow (1) praise, substantive support, or representation of the perpetrator(s) of such attacks; (2) perpetrator-generated content relating to such attacks; or (3) third-party imagery depicting the moment of such attacks on visible victims. Finally, praise, 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 Entity**, defined as an organisation or individual that spreads and encourages hate against others based on their protected characteristics. The entity's activities are characterised by at least some of the following behaviours: • Violence, threatening rhetoric, or dangerous forms of harassment targeting people based on their protected characteristics; • Repeated use of hate speech; • Representation of Hate Ideologies or other designated Hate Entities, and/or; • Glorification or substantive support of other designated Hate Entities or Hate Ideologies.

	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. **Multiple-Victim Violence and Serial murders:** • Facebook considers an event to be a multiple-victim violence or attempted multiple-victim violence 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 perpetrator or an attempted perpetrator of multiple-victim violence. • Facebook considers any individual who has committed two or more murders over multiple incidents or locations a serial murderer. **Hateful Ideologies** • While Facebook's designations of organisations and individuals focus on behaviour, it also recognises that 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 removes 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 removes 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 communities; 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, communities, 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, communities and events will also be removed. Facebook does not allow representation of organisations and individuals designated by Facebook as hate-banned entities. **Militarised social movements (MSMs)**, which include: • **Militia communities**, 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

	○ Coordinate militarised tactical coordination in a present or future armed civil conflict or civil war. • **Communities supporting violent acts amid protests**, defined as non-State actors that repeatedly: ○ 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 • Promotes 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. Furthermore, still in the section on 'Violence and Criminal Behaviour', under its 'Violence and Incitement' policy (Facebook, 2021), Facebook removes language that incites or facilitates serious violence. In particular, Facebook disables accounts and works with law enforcement when it believes there is a genuine risk of physical harm or direct threats to public safety. Users cannot post: ● Threats that could lead to death (and other forms of high-severity violence) and admission of past violence

	targeting people or places where threat is defined as any of the following: ○ Statements of intent to commit high-severity violence, including content where a symbol represents the target and/or includes a visual of an armament or method to represent 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 or method that represents violence. ○ Statements advocating for high-severity violence. ○ Aspirational or conditional statements to commit high-severity violence. ○ Statements admitting to committing high-severity violence except when shared in a context of redemption, self-defence or when committed by law enforcement, military or state security personnel. ● Content that asks or offers services for hire of high-severity violence (for example, hitmen, mercenaries, assassins, female genital mutilation) or advocates for the use of these services. ● Admissions, statements of intent or advocacy, calls to action or aspirational or conditional statements to kidnap or abduct a target or that promotes, supports or advocates for kidnapping or abduction. ● Content that depicts kidnappings or abductions if it is clear the content is not being shared by a victim or their family as a plea for help, or shared for informational, condemnation or awareness raising purposes. ● Threats of high-severity violence using digitally-produced or altered imagery to target living people with armaments, methods of violence or dismemberment. ● Threats that lead to serious injury (mid-severity violence) and admission of past violence towards private individuals, unnamed specified persons, minor public figures, high-risk persons or high-risk groups, where threat is defined as any of the following:

	○ Statements of intent to commit violence, or ○ 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 ○ Statements admitting to committing mid-severity violence except when shared in a context of redemption, self-defence, fight-sports context or when committed by law enforcement, military or state security personnel. ● Content about other target(s) apart from private individuals, minor public figures, high-risk persons or high-risk groups and any credible: ○ Statements of intent to commit violence, or ○ Calls for action of violence, or ○ 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 ● 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 when shared in a context of recreational self-defence, for military training purposes,

	commercial video games, or news coverage (posted by Page or with a 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, conditional or aspirational statements, or advocating for violence due to voting, voter registration or the administration or outcome of an election ● Statements of intent or advocacy, calls to action, or aspirational or conditional statements to bring or take up armaments to locations (including but not limited to places of worship, educational facilities, polling places or locations to count votes or administer an election) or locations where there are temporary 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 7 days. This includes a visual of an armament or method that represents violence that targets these locations. Under this policy, Facebook also prohibits the posting of: • 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. • 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) o 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. o The subject of the threat reports the content to us. • Threats against election workers, including claims of election-related wrongdoing against private individuals when combined with a signal of violence or additional context that confirms that the claim could lead to imminent violence or physical harm. • 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. Finally, in the section titled 'Objectionable content', Facebook Community Standards' feature a 'Violent and Graphic Content' policy (Facebook, 2023) which prohibits particularly violent and sadistic content. Facebook specifies that graphic content can be

	allowed (with limitations such as warning labels or age restrictions) in the context of discussions about important issues such as human rights abuses, armed conflicts or acts of terrorism, to help people condemn and raise awareness. The same section also includes a 'Hate Speech' policy (Facebook, 2023) which prohibits direct attack against people on the basis of their protected characteristics (race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease). Facebook defines attacks as violent or dehumanising speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation. It also prohibits the use of harmful stereotypes, which are defined as dehumanising comparisons that have historically been used to attack, intimidate, or exclude specific groups, and that are often linked with offline violence. Both policies provide examples of prohibited content.
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/. Since the Christchurch terrorist attacks in 2019, 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 offence. 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. Facebook also retrained its AI video detection systems to block violent live streams. By feeding them a dataset of harmful content, Facebook claimed to have reduced the detection time from five minutes to 12 seconds (Sebbagh, 2021).
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 it violates its Community Standards. The Community Standards list the different steps of a standard 'Takedown experience' (Facebook, 2023).
4.1 Notifications of removals or other enforcement decisions	Facebook notifies the user, specifying which policy was violated when possible, so they can understand why the content was

	removed and how to avoid posting violating content in the future.
4.2 Appeal processes against removals or other enforcement decisions	When a user is notified that their content has been removed from the platform for violating the Community Standards, he or she can either accept the decision or disagree with it. If the user accepts the decision, the content remains off Facebook. If the user disagrees with the decision, he or she is invited to specify the reason and the content is then submitted for review. It remains invisible to other users while under review. The assigned reviewer does not know that the post has been reviewed previously. Once the content has been reviewed, the user receives a notification informing them whether the content is back on Facebook or not (Facebook, 2023). In the case where the review confirms the content removal and the user still does not agree, he or she can appeal to Meta's Oversight Board. Composed of a diverse set of experts, this body became operational in 2020 and is meant to independently review some of the most difficult and significant content decisions. Users can appeal a decision about their own content, or about content by others that they have reported. Not all content decisions are eligible for appeal, and the board only selects a certain number of those (Facebook, 2023). The board's decisions are binding, and it also has the option of making non-binding policy recommendations. For instance, in September 2022, the Oversight Board overturned Meta's original decision to remove a Facebook post from a news outlet page reporting a positive announcement from the Taliban regime in Afghanistan on women and girls' education. The board found that removing the post was inconsistent with Facebook's Dangerous Individuals and Organisations policy, which permits reporting on terrorist groups, as well as with its human rights responsibilities (Oversight Board, 2022).
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 Dangerous Individuals and Organisations, through a combination of technology, reports from users and reviews by its teams. Facebook explains that it uses technology 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). Facebook states that its technology finds more than 90% of the content that it removes before anyone reports it, for most violation categories (Facebook, 2022). 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, 2022). With a view to reduce the number of "false positives" (erroneous removal of non-violating content), Facebook implemented a cross-check system to identify content that presents a greater risk of false positives and provide additional levels of review to mitigate that risk. It includes for example entities and posts from journalists reporting from conflict zones and community leaders raising awareness of instances of hate or violence. If during cross-check a reviewer confirms that content violates Facebook's Community Standards, the violating content is addressed accordingly. Depending on the complexity of the content, Facebook may apply multiple levels of review, including in rare instances review by leadership (Meta, 2023). 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. Lastly, Facebook relies on user admins and moderators for moderating content within Facebook groups. Moderators can approve or deny membership requests, approve or deny posts, remove posts and comments, remove and ban people, pin or unpin a post. In addition to these prerogatives, admins can also make another member an admin or moderator, remove an admin or moderator, and manage group settings (Facebook, 2023).

	Facebook is a founding member of GIFCT. In 2022, Meta launched Hasher-Matcher-Actioner (HMA), a free open-source software tool to help platforms identify copies of images or videos and take action against them. Meta is also the chair of the GIFCT for the year 2023.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	If content goes against the Facebook Community Standards, Facebook will remove it and notify the user. Then, Facebook may apply a strike to the account, depending on the severity of the content, the context in which it was shared and when it was posted (Facebook, 2023). A strike can also be applied to a Page or group. Users can see their history of violations, restrictions, and associated duration in their account. In 2023, Facebook updated its strike system, however, the strike system does not apply to the "Dangerous Organisations and Individuals" policy (which includes TVEC). For most violations on Facebook, strikes will lead to the following restrictions: • **One strike:** Warning and no further restrictions. • **2 to 6 strikes:** Restriction from specific features, like posting in groups, commenting, creating a Page, and more. • **7 strikes:** one-day restriction from creating content, which includes posting, commenting, creating a Page, and more. • **8 strikes:** 3-day restriction from creating content. • **9 strikes:** 7-day restriction from creating content. • **10 or more strikes:** 30-day restriction from creating content. Strikes will not be counted on violating content posted over 90 days ago for most violations, or over 4 years ago for the more severe ones. All strikes on Facebook expire after one year. If content posted goes against Facebook's more severe policies, such as the policies on 'Dangerous Individuals and Organisations' or 'Adult Sexual Exploitation', the 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 and using Facebook Live for set periods of time, starting on the first strike. Moreover, Pages and groups that repeatedly violate Facebook's policies may be removed from recommendations and have their

	distribution reduced. Pages may also be restricted from certain monetisation features, and groups may be required to have the admin approve posts. Besides, in a context of civil unrest, Facebook may also restrict accounts by public figures for longer periods of time when they incite or praise ongoing violence. Public figures encompass state and national level government officials, political candidates for those offices, people with over one million fans or followers on social media and people who receive substantial news coverage. In this case, the restriction period is determined on the basis of the severity of the violation and the account's history, the public figure's potential influence, and the severity of the violence. After repeated violations, despite warnings and restrictions or for certain very severe violations, Facebook may also disable a user's account. In some cases, Facebook disables accounts as soon as it becomes aware of them, such as those of dangerous individuals and organisations (Facebook, 2023). Facebook will also remove Pages and groups in the following cases: • If the name, description or cover photo of a Page or group violates our Community Standards. • If an admin of a Page or group creates content, such as posts, comments or rooms, that violates our Community Standards. • If a group moderator creates content that violates our Community Standards. • If a group admin or moderator approves violating content from a group member (Facebook, 2022).
7. Does the service issue transparency reports (TRs) specifically on content related to terrorism and/or violent extremism?	Yes (Facebook, 2023). Facebook issues transparency reports ('Community Standards Enforcement Reports'). The latest available report for Facebook covers Q2 2023. The relevant sections for TVEC are 'Dangerous Organisations: Terrorism and Organised Hate', as well as 'Violence and Graphic Content', 'Violence and Incitement', and 'Hate Speech'. In addition to that, Facebook published this year its first transparency report under the EU Regulation 2021/784 addressing the dissemination of terrorist content online (TCO) (Facebook, 2023), covering the period from 1 June to 31 December 2022. It provides information on removals, blocking, appeals, and administrative or judicial review proceedings, specifically in the EU jurisdiction.
8. What information/fields of data are included in the TRs?	The latest transparency report on Facebook's enforcement of its Community Standards, covering Q2 2023, includes the below-mentioned five fields of information for each category of

	violating content. Additionally, for the category 'Dangerous Organisations: Terrorism and Organised Hate', metrics are broken down into 'Terrorism' and 'Organised Hate'. The report does not include data on other dangerous organisations prohibited from having a presence on Facebook, including those engaging in mass or multiple murder, human trafficking or organised criminal activity: • *Prevalence* *(How prevalent were dangerous organisations violations on Facebook?)* The prevalence metric is the percentage of views that included violating content, for example, terrorism and organised hate. 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 2023, 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 2023, 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 Facebook, it is estimated that no more than five 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 dangerous organisations 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, 2022). 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. 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 and action before people*

	reported it?) This metric shows the percentage of content and accounts actioned for dangerous organisations 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 for dangerous organisations Facebook actioned did people appeal?)* This metric counts the number of pieces of content actioned which were submitted for another review during the reporting period. • *Restored Content (How much actioned content for dangerous organisations was later restored?)* 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 terrorism and organised hate. Its last transparency report notes that: • Content actioned for terrorism increased from 923,000 pieces of content in Q1 2023 to 1.1 million in Q2 2023, due to viral contents that violated its policies. • Appealed content increased from 544,000 in Q1 2023 to 662,000 in Q2 2023, due to an increase in enforcement on non-violating content due to a bug in Meta's proactive detection technology that was later fixed and the content was restored. • Restored content increased from 419,000 in Q1 2023 to 952,000 in Q2 2023 after Meta resolved incorrect actions taken by its proactive detection technology on non-violating content in April.
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 terrorism 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 policy on terrorism. Facebook elaborates on the prevalence methodology in 'Prevalence' (Facebook, 2022). • *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 Community Standards. 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 counts 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 counts 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 (Facebook, 2022). • *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 (Facebook, 2023).
	
	• *Appealed Content.* This metric counts the number of pieces of content (such as posts, photos, videos or comments) 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 (Facebook, 2022).
	
	• *Restored content.* This metric counts the number of pieces of content that Facebook restored during the reporting period after originally taking action on them. 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 (Facebook, 2022).
	
	Updates on the methodologies used by Facebook to measure the metrics included in its transparency reports are available at https://transparency.fb.com/policies/improving/corrections-adjustments/
	
	Lastly, Meta undertook an independent third-party assessment of its Community Standards Enforcement Reports for Facebook

	and Instagram. The assessment was conducted by EY for the period 1 October 2021 to 31 December 2021 and concluded that the calculation of the metrics reported had been prepared based on the specified criteria, were fairly stated, and that Meta's internal controls were suitably designed and operating effectively (Sarang, 2022).
10. Frequency/timing with which TRs are issued	On a quarterly basis.
11. Has this service been used to post TVEC?	Yes. See above sections 7-9. For instance, in 2020, a Facebook internal memo showed how the platform was used to spread TVEC by terrorist and extremist groups across the Middle East (Scott, Facebook did little to moderate posts in the world's most violent countries, 2021).

12. Main changes since last Report	• Facebook updated its Community Standards over the last year, notably its policies on 'Dangerous Organisations and Individuals' in April 2023, and on 'Violence and Incitement' in January 2023. The updates include linguistic precisions, additional examples and changes in the categories and definitions used.
	  ○ The 'Hate Organisations' category was replaced by "Hate Entity" and updated its definition. It now has a broader scope and encompasses individuals (not just organisations).
	  ○ The 'Violence and Incitement' policy provides additional examples of prohibited content and behaviour.
	• Facebook transparency report for Q2 2022 shows a spike in content actioned due to viral content that violated its policy. Portion of this increase was also due to an increase in enforcement on non-violating content due to a bug in Meta's proactive detection technology that was later fixed, and the content was restored. As a result, the report also shows an increase in the number of appeals and restored content. Both Facebook and Instagram were affected.
	• Facebook issued its first transparency report under the EU Regulation 2021/794 addressing the dissemination of terrorist content online (TCO), covering Q2 2022.
	• Since 2020, after an unsuccessful appeal regarding a content moderation decision, users can submit an appeal to Meta's Oversight Board for a second review, provided that their case is eligible and gets selected.

	• Meta undertook an independent third-party assessment of its Community Standards Enforcement Reports for Facebook and Instagram. The assessment was conducted by EY for the period 1 October 2021 to 31 December 2021 and concluded that the calculation of the metrics reported had been prepared based on the specified criteria, were fairly stated, and that Meta's internal controls were suitably designed and operating effectively.

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 (YouTube/Google, 2023), under the section titled 'Violent or dangerous content', contain several sub-categories of violating content that are relevant to terrorist and violent extremist content. The 'Violent extremist or criminal organisations policy' states that content intended to praise, promote, or aid violent extremist criminal organisations is not allowed on YouTube for any purpose, including recruitment. YouTube prohibits the following types of content: • Content produced by violent extremist, criminal or terrorist organisations • Content praising or memorialising prominent terrorist, extremist, or criminal figures in order to encourage others to carry out acts of violence • Content praising or justifying violent acts carried out by violent extremist, criminal, or terrorist organisations • Content aimed at recruiting new members to violent extremist, criminal, or terrorist organisations • Content depicting hostages or posted with the intent to solicit, threaten, or intimidate on behalf of a violent extremist, criminal, or terrorist organisation • Content that depicts the insignia, logos, or symbols of violent extremist, criminal, or terrorist organisations in order to praise or promote them • Content that glorifies or promotes violent tragedies, such as school shootings

	YouTube relies on many factors, including government and international organisation designations, to determine what constitutes criminal or terrorist organisations. For example, it terminates any channel where there is reasonable belief that the account holder is a member of a designated terrorist organisation, such as a Foreign Terrorist Organisation (US), or organisations identified by the United Nations. 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. Graphic or controversial footage with sufficient content may be subjected to age restrictions or a warning screen. The 'Violent extremist or criminal organisations policy' also gives the following examples of content that is not allowed on YouTube: • Raw and unmodified reuploads of content created by terrorist, criminal, or extremist 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 • Footage filmed by the perpetrator during a deadly or major violent event, in which weapons, violence, or injured victims are visible or audible • Video game content which has been developed or modified ('modded') to glorify a violent event, its perpetrators, or support violent criminal or terrorist organisations • Glorifying violence against civilians • Fundraising for violent criminal, extremist, or terrorist organisations Moreover, YouTube's 'Violent or graphic content policies' prohibit 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 • Fights involving minors • Footage, audio, or imagery involving road accidents, natural disasters, war aftermath, terrorist attack aftermath, street fights, physical attacks, immolation, torture, corpses, protests or riots, robberies, medical procedures or other such scenarios with the intent to shock or disgust viewers • Dramatized or fictional footage of content prohibited by these guidelines where the viewer is not given enough context to understand that the footage is dramatized or fictional • Footage filmed by the perpetrator during a deadly or major violent event, in which weapons, violence, or injured victims are visible or audible (even if there is educational, documentary, scientific, or artistic context provided). In turn, YouTube's 'Hate speech policy' bans content promoting violence or hatred against individuals or groups based on any of the following attributes: Age, Caste, Disability, Ethnicity, Gender Identity and Expression, 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. 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). Furthermore, 'Harmful or dangerous content policies' ban 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.

	Lastly, Google's Autocomplete policies (Google, 2023), which apply to YouTube, prohibit autofill predictions that are terrorist content, hateful content, dangerous content, or violence and gore.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	YouTube's Community Guidelines are available at https://support.google.com/youtube/answer/9288567?hl=en
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No. YouTube's Community Guidelines apply to all types of content available on the platform, including videos, video descriptions, comments, live streams, unlisted and private content, comments, links, thumbnails, Community posts, thumbnails, and any other YouTube product or feature.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	If content violates any of YouTube's content policies, YouTube removes the content.
4.1 Notifications of removals or other enforcement decisions	The content removal is notified to users via email, desktop or mobile notifications, and an alert in their channel settings. 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, 2023). 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 offence 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, 2023) turned on. If the content is a custom thumbnail, it will be removed. • If YouTube finds that the content was in violation of the 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. When a video is removed without a strike or a warning, the user who posted it can also appeal the decision. 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. Data inputs are used to train automated systems to find patterns. Automated systems use those patterns to make predictions about new examples to match and automatically remove content that they are highly confident violates policies, before it is viewed, and flag the rest for human review. Content reviewers then evaluate the flagged content to determine whether or not they violate any of YouTube's Community Guidelines and take action accordingly, while protecting content that has an educational, documentary, scientific or artistic purpose. Reviewers' inputs are used to train and improve the accuracy of decisions made by algorithmic systems on a much larger scale (YouTube/Google, 2023). 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, 2023). YouTube also relies on its community to flag inappropriate content, through a user report mechanism. In particular, users can report inappropriate videos, channels and other content on YouTube such as a Short, a playlist, a thumbnail, a link in a video's description, a comment, a live chat message, or an ad. In addition, users can report a search prediction, if they believe is violates YouTube's Autocomplete policies, which prohibit terrorist content, dangerous content, hateful content, and violence and gore.

	In June 2022, YouTube rebranded its ten-year old “Trusted Flagger” programme’ to the “Priority Flagger” programme, which provides policy training and tools for government agencies and NGOs who have expertise in a specific policy area relating to YouTube’s Community Guidelines. This programme is not open to individual users. As of May 2022, only government agencies and NGOs are eligible for participation in the YouTube Priority Flagger programme. Participants must have identified expertise in at least one policy area, flag content frequently with high accurate rate, and be open to ongoing discussion and feedback on various YouTube content areas. Certain organisations, including organisations from countries or regions where there is a history of human rights abuses or speech suppression, may be subject to further review (Google/YouTube, 2023). This change comes at a time when YouTube is increasingly relying on automated tools for content moderation (Hale, 2022). Individuals can apply to the YouTube Contributors Program (YouTube/Google, 2023), through which they can answer questions from users and produce help videos (YouTube/Google, 2023). Content reported by Priority 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 Priority Flaggers are prioritised for review by YouTube’s teams (Google/YouTube, 2023). With respect to the automated systems that detect extremist content, YouTube’s teams have manually reviewed over three million videos to provide large volumes of training examples, which help improve the machine-learning flagging technology. Between April and June 2022, approximately 95% of the videos that were removed for violating YouTube’s ‘Violent Extremism policy’ were first automatically flagged (Google/YouTube, 2023). YouTube continues to invest in a network of over 300 government partners and NGOs who bring expertise to the platform’s enforcement systems, including through YouTube’s Priority Flagger programme (Google/YouTube, 2023). YouTube is a founding member of GIFCT. YouTube contributed over 45,000 unique hashes to the hash-sharing database in 2023.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	In case of violations, YouTube applies the following strike system: **Warning**

	The first time a user posts content that violates YouTube's Community Guidelines, they will likely receive a warning. In August 2023, YouTube launched an updated approach to the warning system, giving creators the option of taking an educational training course when they receive a Community Guidelines warning. Completion of the course will lift the warning from a creator's channel — so long as they do not violate the same policy for 90 days. For subsequent violations, YouTube issues a 'strike' against the user's channel. The channel is terminated if the user receives three strikes within a 90-day period. However, sometimes a single case of severe abuse will result in channel termination without warning. **First strike** When the first strike is issued, the user cannot do any of the following for one week: • Upload videos, live streams, or stories • Start a scheduled live stream • Schedule a video to become public • Create a Premiere • Add a trailer to an upcoming Premiere or live stream • 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 The user's scheduled public content is set to 'private' for the penalty duration. Full privileges are restored automatically after the 1-week period, but the strike will remain on the user's channel for 90 days. **Second strike** 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. **Third strike**

	Three strikes in the same 90-day period will result in the user's channel being permanently removed from YouTube (Google/YouTube, 2023). Beyond the three strikes system, YouTube will terminate a channel or an account for the following reasons: • Repeated violations of the Community Guidelines or Terms of Service across any form of content • A single case of severe abuse (such as predatory behaviour, spam, or pornography) • Dedication to a policy violation (like hate speech, harassment, or impersonation) 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, 2023).
7. Does the service issue transparency reports (TRs) on TVEC?	Yes (Google, n.d.). YouTube issues transparency reports on the enforcement of its Community Guidelines. One section of these reports is about 'Violent Extremism' (Google/YouTube, 2023). The last transparency report, which covers the period going from July to September 2023, 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: **Removals**

	• Number of channels removed, broken down by removal reason among which the promotion of violence and violent extremism • Number of videos removed, broken down by source of first detection (automated flagging, user, organisation, government agency) • Percentage videos removed, broken down by number of views (0 view, 1 to 10 views, and more than 10 views) • Number and percentage of videos removed, broken down by removal reason among which the promotion of violence and violent extremism • Number of videos removed by country/region • Number of comments removed • Percentage of comments removed by source of first detection (automated flagging and human flagging) • Number and percentage of comments removed, broken down by removal reason among which the promotion of violence and violent extremism **Views** • 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). **Appeals** • Total number of appeals that YouTube received for videos removed due to a community violation per quarter • Total number of videos reinstated **Flags** • Number of videos flagged by all human flaggers • Percentage of flags, broken down per type of human flaggers (user, organisation, other) • Top 10 countries/regions by human flagging volume • Number and percentage of human flags, broken down by flagging reason among which 'promotes terrorism' **Government requests**

	• Number of removal requests received since 2011, broken down by reasons, products, and requesters • Number of items named for removal since 2011 **Featured policies** YouTube's transparency report features a section titled 'Featured policies', which includes 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 uses 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, 2023).
10. Frequency/timing with which TRs are issued	On a quarterly basis.
11. Has this service been used to post TVEC?	Yes. See above sections 7-8.
12. Main changes since last Report	• YouTube updated its Community Guidelines and modified its categories of violating content. The 'Violent criminal organisations policy' became the 'Violent extremist and criminal organisations policy' with a new focus on 'violent extremism'. The policy is more precise and contains additional examples. The Guidelines also now refer to designation lists of terrorist organisations such as the Foreign Terrorist Organisation list (US) and the United Nations'. • In June 2022, YouTube rebranded its ten-year old "Trusted Flagger" programme' to the "Priority Flagger" programme, which provides policy training and tools for government agencies and NGOs who have expertise in a specific policy area relating to YouTube's Community Guidelines. This programme is not open to individual users. As of May 2022, only government agencies and NGOs are eligible for participation in the YouTube Priority Flagger programme. This change comes at a time when YouTube is increasingly relying on automated tools for content moderation.

	• YouTube also updated its warning/strike system which now has optional policy training to correct policy violations and additional penalties for channels that have received a strike. For example, after one strike, users may not start a scheduled live stream or schedule a video to become public for one week. YouTube also added that a single case of severe abuse can sometimes result in channel termination without warning.

3. 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. Instagram also prohibits content that contains credible threats or hate speech, and content that targets private individuals to degrade or shame them. 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 (removal of content). In particular: are there notifications of removals	Instagram removes or blocks content from the platform when content violates its Terms of Service or Community Guidelines, or if it is required to do so by law. Instagram can also stop providing all or part of the service to a user.

or other enforcement decisions and appeal processes against them?	
4.1 Notifications of removals or other enforcement decisions	If content goes against Instagram's Terms of Service or 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 appeal 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,
6. Sanctions/consequences in case of breaches of ToS or Community Guidelines/Standards	If content goes against Instagram's Terms of Service or 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, 2023). Instagram follows the same approach as Facebook regarding sanctions (Facebook, 2023). However, the strike system does not apply to the "Dangerous Organisations and Individuals" policy (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, 2023) ('Community Standards Enforcement Reports'). The latest available report for Instagram covers Q2 2023. The relevant sections for TVEC are "Dangerous Organisations: Terrorism and Organised Hate", as well as "Violence and Graphic Content", "Violence and Incitement", and "Hate Speech". In addition to that, Instagram published this year its first transparency report under the EU Regulation 2021/784 addressing the dissemination of terrorist content online (TCO) (Instagram, 2023), covering the period from 1 June to 31 December 2022. It provides information on removals, blocking, appeals, and administrative or judicial review proceedings,

	specifically in the EU jurisdiction. Under the TCO, such reports have to be published annually.
8. What information/fields of data are included in the TRs?	The latest transparency report on Instagram's enforcement of its Community Standards, covering Q2 2023, includes the below-mentioned five fields of information for each category of violating content. Additionally, for the category 'Dangerous Organisations: Terrorism and Organised Hate', metrics are broken down into 'Terrorism' and 'Organised Hate'. The report does not include data on other dangerous organisations prohibited from having a presence on Instagram, including those engaging in mass or multiple murder, human trafficking or organised criminal activity: • *Prevalence* *(How prevalent were dangerous organisations violations on Instagram?)* The prevalence metric is the percentage of views that included violating content, for example, terrorism and organised hate. 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 2023, 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 2023, the upper limit was 0.05% for violations of the policy for terrorism on Instagram. This means that out of every 10,000 views of content on Instagram, it is estimated that no more than five 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 dangerous organisations content did Instagram take action on?)* Facebook indicates that a piece of content can be 'any number of things', including a post, photo, video or comment (Facebook, 2022). 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. This includes content that

	Instagram actioned on after someone reported, and content Facebook found proactively. • *Proactive rate (Of the violating content actioned, how much did Instagram find and action before people reported it?)* This metric shows the percentage of content and accounts actioned for dangerous organisations 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 for dangerous organisations Instagram actioned did people appeal?)* This metric counts the number of pieces of content actioned which were submitted for another review during the reporting period. • *Restored Content (How much actioned content for dangerous organisations was later restored?)* 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 terrorism and organised hate. Its last transparency report notes that: • Content actioned for terrorism increased from 1.6 million pieces of content in Q1 2023 to 2 million in Q2 2023, due to viral contents that violated its policies. Portion of the increase was also due to an increase in enforcement on non-violating content due to a bug in Meta's proactive detection technology that was later fixed and the content was restored. • Proactive rate on Terrorism increased from 95.1% in Q1 2023 to 98.3% in Q2 2023, due to an increased action in proactive detection technology. • Appealed content for terrorism increased from 81,700 in Q1 2023 to 144,000 in Q2 2023, due to an increase in enforcement on non-violating content due to a bug in Meta's proactive detection technology that was later fixed and the content was restored. • Restored content for terrorism increased from 14,400 in Q1 2023 to 643,000 in Q2 2023 after Meta resolved incorrect actions taken by its proactive detection technology on non-violating content in April. • Content actioned for organised hate decreased from 408,000 in Q1 2023 to 215,000 in Q2 2023, returning to

	pre-Q1 levels after a spike in reported viral content in February. • Proactive rate on organised hate decreased from 91.6% in Q1 2023 to 84.5% in Q2 2023 due to continued updates made to Meta proactive detection technology to improve accuracy, leading to less automated enforcement.
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, 2022). • *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. Lastly, Meta undertook an independent third-party assessment of its Community Standards Enforcement Reports for Facebook and Instagram. The assessment was conducted by EY for the period 1 October 2021 to 31 December 2021 and concluded that the calculation of the metrics reported had been prepared based on the specified criteria, were fairly stated, and that Meta's internal controls were suitably designed and operating effectively (Sarang, 2022).
10. Frequency/timing with which TRs are issued	On a quarterly basis, jointly with Facebook. However, every time a new quarterly report is published, previous data disappears from the website.
11. Has this service been used to post TVEC?	Yes. The media has covered many examples: (Carmen, 2015) (Hymas, 2019) (Cox, 2019).
12. Main changes since last Report	• Instagram transparency report for Q2 2022 shows a spike in content actioned due to viral content that violated its policy. Portion of this increase was also due to an increase in enforcement on non-violating content due to a bug in Meta's proactive detection technology that was later fixed and the content was restored. As a result, the report also shows an increase in the number of appeals and restored content. Both Facebook and Instagram were affected. • Instagram issued its first transparency report under the EU Regulation 2021/794 addressing the dissemination of terrorist content online (TCO), covering Q2 2022. • Since 2020, after an unsuccessful appeal regarding a content moderation decision, users can submit an appeal to Meta's Oversight Board for a second review, provided that their case is eligible and gets selected. • Meta undertook an independent third-party assessment of its Community Standards Enforcement Reports for Facebook and Instagram. The assessment was conducted by EY for the period 1 October 2021 to 31 December 2021 and concluded that the calculation of the metrics reported had been prepared based on the specified criteria, were fairly stated, and that Meta's internal controls were suitably designed and operating effectively.

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, Security, and Integrity' the ToS state that WhatsApp works to protect the safety, security, and integrity 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. There is no definition of what constitutes "abusive people and activity'". 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).
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 a livestream feature.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	WhatsApp broadly states that it may modify, suspend, or terminate a user's access to or use of its services at any time for any reason, such as if the user violates "the letter or spirit of [its] Terms or create harm, risk, or possible legal exposure for [WhatsApp], its users, or others".
4.1 Notifications of removals or other enforcement decisions	If a number is banned, the user sees a message when he or she opens the app (WhatsApp, 2023).
4.2 Appeal processes against removals or other enforcement decisions	If a user believes that their account was terminated or suspended by mistake, the user can contact WhatsApp via email or request a review in the app. WhatsApp indicates

	that it "may not issue a warning before banning [an] account" (WhatsApp, 2023).
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, for example a contact, other users' status, and channels in countries where this feature is available. If a user reports a contact, WhatsApp receives the last five messages sent to you by the reported user or group, without notifying them (WhatsApp, 2023). Then, WhatsApp's moderators review those reports to take appropriate action. 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. In addition to responding to and actioning on user reports, WhatsApp explains that it deploys tools and resources to prevent harmful behaviour on the platform, focusing on prevention (WhatsApp, 2023).
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. Since 2021, WhatsApp issues monthly reports specifically concerning the Indian jurisdiction, under the new Information Technology Rules. These reports contain information on actions taken by WhatsApp in responses to grievances received from users, and accounts actioned through the prevention and detection methods for violating the laws of India or WhatsApp's TS. The reports feature the total number of user reports received, and number of accounts actioned. Both are broken down per topic of user complaint, among

	which a 'Safety' category, broadly defined as "issues that may be about abuse of harmful behaviour on the platform" (WhatsApp, 2023). For instance, in May 2023, in India, WhatsApp received 37 reports for "Safety" and actioned one account. It also banned over 6 million problematic accounts.
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. After the Christchurch shootings, two far-right violent extremists were reportedly part of a WhatsApp group called 'Christian White Militia' and published statements encouraging terrorism in March 2019 (Dearden, 2019). In the 2017 Westminster terrorist attack, the perpetrator used WhatsApp a few minutes before the attack, notably to communicate and diffuse jihadist propaganda (Hill QC, February 2018).
12. Main changes since last Report	No main changes since last Report.

5. Weixin/WeChat

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	WeChat does not define TVEC. However, WeChat updated its Community Guidelines (WeChat, 2023) which are much more detailed and define 'terrorism', 'violent extremism', and 'hate speech'. In the section titled 'Terrorism, Violent Extremism and Other Criminal Behaviour', WeChat prohibits any content or behaviour that amounts to criminal behaviour. This includes any content or conduct that may constitute a genuine risk or harm or direct threat to public safety, or any content or conduct that breaches any applicable laws or regulations. Terrorism is defined as the "use of political violence and the exploitation of fear aimed at reaching at a target audience, with the end goal of bringing behavioural and societal change through coercion, whether that be to further a religious, political or any other ideological cause". Violent extremism is defined as "a form of extremism that condones and enacts violence with ideological or deliberate

	intent, such as religious or political violence." WeChat adds that violent extremist views can manifest in connection with a range of issues, including politics, religion and gender relations. Extremism is a term used to characterise a variety of attitudes, beliefs, and behaviours that often are on the extreme end of the political, religious, or ideological spectrum within society (e.g., white nationalist, anarchist).
	
	WeChat states that it takes a very strict stance against terrorism and violent extremism, and does not allow dangerous individuals or organisations to use the platform to promote terrorism, violent extremism, crime, or other types of harmful activities. If WeChat believes there is a threat to the safety of the platform or users' safety, it may suspend or terminate such users' accounts and notify the relevant legal authorities. As such, content which constitute statements, calls to action and advocacy for the following are prohibited:
	
	• terrorist activity and terrorist organisations;
	• organised hate;
	• mass or serial murder;
	• human trafficking; and
	• organised violence, riots, or criminal activity.
	
	Then, under the section titled 'Violent Content' of its Community Guidelines, WeChat prohibits content that is particularly violent or graphic, such as pictures or videos depicting dismemberment, visible organs, bodies, human torture and animal abuse, to safeguard the mental and physical well-being of our users.
	
	• Incitement: Users may not upload, share, publish, transmit, stream or otherwise make available on or through the WeChat platform any content that:
	  ○ depicts or promotes cruelty, violence or unlawful acts towards persons (including but not limited to playground fights between minors, dismemberment, visible internal organs and human torture);
	  ○ encourages others to commit acts of violence against individuals or a specified group of people; or
	  ○ c) contains content concerning street battles, physical attacks, sexual assaults, torture, bodies or violent demonstrations.
	• Violence and threats: Users may not upload, share, publish, transmit, stream or otherwise make available on or through the WeChat platform any content that contains:

	○ attempts, threats or statements of intent to commit violence which may lead to death or serious injury; ○ admission of past violence targeting people or places; or ○ instructions on how to make or use weapons or explosives with a goal to seriously injure or kill people. WeChat's Community Guidelines, under the section titled 'Hateful, Spam or Other Inappropriate Behaviour', defines hate speech as "a direct attack against people on the basis of certain protected characteristics, including but not limited to age, race, ethnicity, nationality, disability, religion, caste, sexual orientation, sex, gender identity, immigration status and serious disease". This could be by way of violent or dehumanising speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation. Under this policy, users are not allowed to upload, share, publish, transmit, stream or otherwise make available on or through the WeChat platform any content that: • Contains or incites hatred or discrimination against any persons due to any of the protected characteristics set out above; or • is, or could reasonably be interpreted as being, hateful, harassing, abusive (physical, sexual or verbal), racially or ethnically offensive, discriminatory towards a particular class, individual or group, defamatory, humiliating or disrespectful to other people, local ethics or customs (publicly or otherwise), threatening, profane or otherwise objectionable, including dehumanising speech or imagery, mocking of victims of hate crimes and derogatory remarks about physical appearance or intellectual capacity. In addition, in its Acceptable Use Policy (WeChat, 2023), under Section 2 titled 'Violence and Crime', WeChat prohibits violent, criminal, illegal, or inappropriate content, as well as any content which may (in its opinion) constitute a genuine risk of harm or direct threat to public safety, and any content 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. Under Section 5 of its Acceptable Use Policy, titled 'Objectionable Content', WeChat also prohibits any content and 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://safety.wechat.com/en_US/community-guidelines and https://www.wechat.com/en/acceptable_use_policy.html
3. Are there specific provisions applicable to livestreamed content	No.

in 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?	WeChats reserves the right to determine whether content violates its Community Guidelines. In response, it may take enforcement actions ranging from issuing a warning to removing content or terminating the user's account. WeChat also 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 WeChat has completed a review of a user report, it notifies the person reporting whether action has been taken or not. If WeChat decides to take action against a reported user, they will be notified as well.
4.2 Appeal processes against removals or other enforcement decisions	WeChat's content moderation decisions can be appealed by either the reported user after that WeChat has taken action on, or the reporting user after WeChat has decided not to take action against the reported user. Users who disagree with the content moderation decision can fill out a complaint using the feedback form in the WeChat Help Centre.
5. Means of identifying TVEC (for example, monitoring algorithms, user-generated, human (staff) reviewers, hash-sharing/URL sharing database)	WeChat has a user report mechanism in place. The 'report button' is available inside the app. Users can make a report against content in a chat or group chat, or against another WeChat user. Users also have to select the most relevant category of violating content. When the report is made against content, users are also required to provide the messages, and have the option to add image evidence and a written complaint. Then, WeChat's content moderation team will review the user report to determine whether there has been a violation of these Community Guidelines, and take action accordingly. In addition to human review, WeChat also uses artificial intelligence and other content moderation tools to assist its work with the review of user reports and identification of violating content. Besides, it has been reported that Chinese online firms, including WeChat, have a team of moderators policing problematic content[51]. Political activists have declared having been followed based on what they have said on WeChat, and chat records have turned up as evidence in court (Zhong, 2018). Also, research has shown that WeChat uses algorithmic technology (Knockel, Ruan, Crete-Nishihata, & Deibert, 2018), keyword filtering and URL blocking (Ruan L. J.-N., 2016) to

	censor content that is in violation of its ToS (which may include the posting of TVEC). Although these methods had been reportedly applied only to accounts registered to mainland China phone numbers (Ruan L. J.-N., 2016), recent research has shown that international (i.e. non-Chinese) accounts are also monitored 'to invisibly train and build up WeChat's Chinese political censorship system' (Knockel, et al., 2020) WeChat is not a member of the GIFCT.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	When it determines that content violates its Community Guidelines, WeChat may take any or all of the following actions against violating users: • Issue warnings to violating users; • Hide or remove the offending content (or potentially offending content). Please note that if the content has already been received or viewed by the user, we will not be able to hide or delete the content as it is already stored in the user's device cache. However, we may limit further dissemination or visibility of other users who have not received or viewed the relevant content; • Display notifications to users who receive violating content reminding them to take precautions; • Restrict violating users from using certain account functions, suspend or terminate the user's account; • Notify and cooperate with relevant judicial and law enforcement agencies if we have reason to believe that a user has committed a crime or as required by applicable law. WeChat considers a variety of factors when deciding what action to take against a violating user, including but not limited to: • the severity of the breach; • whether the violation was intentional; • the number of users affected by the breach; • whether the violation constitutes an illegal act under applicable law; • whether the user has committed the same or similar violations in the past; • whether there have been user reports by other users for the same or similar violations against the same 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?	Yes. The Christchurch shooting was posted on WeChat (Kenny, 2019). In addition, WeChat has been used to disseminate anti-Muslim propaganda (Huang, 2018).
12. Main changes since last Report	• WeChat released new Community Guidelines that are much more detailed. They provide definitions for 'terrorism', 'violent extremism', and 'hate speech', as well as additional explanations and examples for each category of violating content. • WeChat discloses more information on detection methods, acknowledging the use of AI tools for content moderation. • WeChat discloses more information on enforcement methods: it provides a list of factors considered when taking action against a violating user.

6. 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), in the section titled 'Your submissions to our services', prohibit users from posting objectionable, offensive, unlawful, deceptive or harmful content, such as comments, ratings and reviews, pictures, videos, and podcasts (including associated metadata and artwork); as well as from planning or engaging in any illegal, fraudulent, or manipulative activity.

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 allows 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. iMessage and FaceTime content are protected by end-to-end, which means that only the sender and receiver can access them. Thus, until recently, it was difficult to imagine how an algorithm or an on-staff reviewer who works for Apple could detect any problematic content, such as TVEC. In 2021, Apple launched new 'Communication Safety' features to combat child sexual abuse material (CSAM) (Apple, 2023). 'Communication Safety' scans messages locally on young users' devices to flag content that children send or receive in iMessages that contain nudity. The feature is also expanding to FaceTime, and Apple debuted to develop additional features tailored for adults. For now, it is exclusively used to detect CSAM, and not any other types of illegal content (Newman, 2023). Apple is not a member of the GIFCT.
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 Terms and Conditions, Apple may, without notice to the user, terminate the user's Apple

	ID, license to Apple's software and/or access to its services, which include iMessage and FaceTime.
7. Does the service issue transparency reports (TRs) on TVEC?	No. Apple does issue transparency reports (Apple, n.d.) that contain a section on content removal requests from governments and private parties reporting violations of its Terms and Conditions or local laws, but there is no specific information on TVEC. In 2022, Apple published the first App Store transparency report (Apple, 2022), which includes the number of App submission rejections, broken down by App Store Review Guidelines section. Among these is a 'safety' category, which includes terrorist content.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Possibly. A security manual issued by ISIS recommended use of iMessage to protect supporters' identities, (Zetter, 2015) but there is no evidence that ISIS supporters have actually used it (Dilger, 2015). Also, the FBI recently managed to unlock the iPhone of the perpetrator of the Pensacola attack, finding that he had been in contact with al-Qaeda 'using end-to-end encrypted apps.' However, it is not clear whether iMessage or FaceTime were actually used (Sky News, 2020).
12. Main changes since last Report	No main changes since last Report.

7. 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 (TikTok, 2023) feature detailed explanations of what content and organisations are banned from the platform. The Guidelines were updated in April 2023. **Violent and hateful organisations and individuals:** This category replaces the former category 'Violent extremist organisations and individuals'. It does not specifically include 'terrorist organisations' anymore, whereas before they were

	clearly defined and prohibited. Instead, it now features 'violent political organisations' and 'hateful organisations'. Under this policy: • Violent extremists are non-states groups, including those designated by the United Nations, that threaten or use violence against civilians for political, religious, ethnic, or ideological reasons. • Criminal organisations are transnational, national, or local groups that commit serious crimes, including violence, trafficking, kidnapping, financial crimes, and cybercrime. • Violent political organisations are non-state groups that commit violent acts that primarily target non-civilians and are acting legitimately under a right of self-determination according to international law, such as the United Nations Charter, a United Nations resolution, the International Covenant on Civil and Political Rights (ICCPR), and the international Court of Justice (ICJ). • Hateful organisations are groups who target people based on protected attributes, including inciting hate, dehumanising individuals or groups, and promoting hateful ideologies. Thus, the following content is strictly prohibited: • Accounts operated by organisations or individuals that promote violence or hateful ideologies on or off-platform • Promoting or materially supporting a violent or hateful organisation, including any praise or celebration, or the provision of goods or services • Promoting or materially supporting any violence committed by a violent political organisation • Promoting or materially supporting individuals who are perpetrators of mass violence or who promote hateful ideologies If TikTok becomes aware that any such actor may be on the platform, it will conduct a thorough review – including off-platform behaviour – which may result in an account ban. TikTok does not allow anyone to promote or materially support violent or hateful actors. Content that may appear neutral, such as referencing a quote from a hateful organisation, must make clear that there is no intent to promote it. TikTok may make limited allowances for people to discuss violent political organisations, but only if: (1) their causes are recognised as legitimate under international legal

	frameworks, (2) they do not primarily target civilians, and (3) the content does not mention violence. The following may be allowed: • Discussing a violent political organisation (as long as there is no mention of violence) • Educational and documentary content that raises awareness of the harms caused by violent and hateful actors **Violent behaviours and criminal activities:** This category replaces the former one titled 'Threats and incitement to violence.' Under this policy, TikTok does not allow any violent threats, incitement to violence, or promotion of criminal activities that may harm people, animals, or property. If there is a specific, credible, and imminent threat to human life or serious physical injury, we report it to relevant law enforcement authorities. Thus, the following content is strictly prohibited: • Threatening or expressing a desire to cause physical injury to a person or a group • Promoting or inciting violence, such as making a general call for an attack, encouraging others to attack, and recommending people bring weapons to a location to intimidate others • Promoting any type of theft, or the criminal destruction of property or the natural environment • Providing instructions on how to commit criminal activities that may harm people, animals, or property However, the following may be allowed: • Threats of violence in completely fictional settings (as long as there is no relevance or reference to the real world) **Hate speech and hateful behaviours:** This category regroups the former ones titled 'Hateful behaviour', 'Slurs', and 'Hateful ideologies'. Under this policy, TikTok does not allow any hateful behaviour, hate speech, or promotion of hateful ideologies. This includes content that attacks a person or group because of protected attributes, including: caste, ethnicity, national origin, race, religion, tribe, immigration status, gender, gender Identity, sex, sexual orientation, disability, serious disease.

	Hateful ideologies are systems of beliefs that exclude, oppress, or otherwise discriminate against individuals based on their protected attributes, such as racial supremacy, misogyny, anti-LGBTQIA+, and antisemitism. Protected attributes are personal characteristics that we are born with, are immutable, or cannot change without severe psychological harm, and which may result in disproportionate stigmatization. In addition, we also provide some protections related to age, and may consider other protected attributes when we have additional context, such as specific regional information provided to us by a local non-governmental organisation (NGO). The attributes above are informed by the Universal Declaration of Human Rights and international conventions. The following content is strictly prohibited: • Promoting violence, exclusion, segregation, discrimination, and other harms on the basis of a protected attribute • Promoting any hateful ideology or claiming supremacy over a group of people on the basis of protected attributes • Demeaning someone on the basis of their protected attributes by saying or implying they are physically, mentally, or morally inferior, or calling them degrading terms, such as criminals, animals, and inanimate objects • Using a hateful slur associated with a protected attribute • Denying well-documented historical events that harmed groups based on a protected attribute, such as denial of the Holocaust or the genocide against the Tutsi in Rwanda • Promoting or advertising conversion therapy or related programs that attempt to change a person's sexual orientation or gender identity • Intentionally targeting people who are transgender or gender non-conforming by referring to them using their former name or gender rather than their current name or expressed gender (deadnaming or misgendering) • Facilitating the trade of any items that promote hate speech or hateful ideologies, such as books and clothing with hateful logos However, the following may be allowed:

	• Self-referential slurs used by a member of a group with that particular protected attribute • Educational and documentary content raising awareness against hate speech **Shocking and graphic content:** TikTok does not allow gory, gruesome, disturbing, or extremely violent content. Content is age-restricted if it shows human or animal blood. Content is ineligible for the For You feed (FYF) if it shows fictional violence, blood, potentially distressing or mildly graphic material, animal genitalia or sexual activity between animals. Content is also ineligible for the FYF if it contains graphic footage of events that are in the public interest to view. The following content is strictly prohibited: • Real-world torture, graphic violence, and extreme physical fighting • Graphic deaths and accidents • Body parts that are dismembered, mutilated, charred, burned, or severely injured The following content is age-restricted and may only be viewed by accounts 18 years or older: • Blood of humans and animals The following content is ineligible for the "For You" feed: • Graphic or potentially distressing footage of events in the public interest to view, such as clashes with law enforcement and the aftermath of a bombing or natural disaster • Fictional graphic violence and extreme physical fighting • Blood of humans and animals • Potentially distressing material that may cause anxiety or fear, such as showing non-severe injuries and accidents, dead animals, jump scare effects, and gory make-up • Mildly graphic material that may cause disgust, such as human and animal bodily functions and fluids (such as urine or vomit), and close-ups of organs and certain animals (such as insects, rats) • Genitalia and sexual activity of animals The following content may be allowed:

	• Professional fighting, such as boxing and mixed martial arts • Blood shown in an educational context (such as menstruation) and artistic settings (such as fine art) • Food-related blood products, such as blood sausage, blood/black pudding, curd, and cake
2. Manner in which the ToS or Community Guidelines/Standards are communicated	TikTok's Community Guidelines are available at https://www.tiktok.com/community-guidelines/en/overview/ TikTok's Terms of Service are available at • US: https://www.tiktok.com/legal/page/us/terms-of-service/en • EEA/UK/CH: https://www.tiktok.com/legal/page/eea/terms-of-service/en • Rest of world: https://www.tiktok.com/legal/page/row/ TikTok's Community Guidelines and Terms of Service are available within the application and online. TikTok users are also notified of any updates to the Terms of Service or Community Guidelines.
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	Yes. TikTok's Community Guidelines apply to livestreamed content. Community members must be 18 years and older to go live and to send gifts to a creator during a live session. Violations of Community Guidelines during a livestream will result in closing an ongoing livestream session, and may lead to restrictions on using live-streaming or an account ban. Repeatedly live-streaming content that is ineligible for the "For You" feed may result in the individual's account not being eligible for the "For You" feed or being harder to find in 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?	TikTok states that its approach to content moderation relies on four pillars: • Remove violative content from the platform that breaks its rules • Age-restrict mature content so it is only viewed by adults (18 years and older) • Maintain For You feed (FYF) eligibility standards to help ensure any content that may be promoted by its recommendation system is appropriate for a broad audience • Empower its community with information, tools, and resources

	Consistent with its Community Guidelines, TikTok removes content - including video, audio, livestream, images, comments, and text - that violates those guidelines. The Community Guidelines are informed by international legal frameworks, industry best practices, and input from TikTok's community, safety and public health experts, and TikTok's regional Advisory Councils. TikTok regularly reviews and updates its Community Guidelines to evolve alongside new behaviours and risks with the goal of creating a safe and entertaining experience for its diverse community (TikTok, 2023).
4.1 Notifications of removals or other enforcement decisions	Yes. If a member of TikTok's community has posted a video, comment, audio, or livestreamed content that is not allowed, they will be notified in the app along with the violation reason. If the account has been banned, they will receive a banner notification when they next open the app, informing them of this.
4.2 Appeal processes against removals or other enforcement decisions	TikTok offers creators the ability to appeal an enforcement decision (content or account removal). TikTok reviews it and, if the appeal is approved, the content or account is reinstated. The strike is removed from the user's account (TikTok, 2023).

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 uses computer vision models to help detect visual signals, emblems, and logos that are known to be associated with extremist and hate groups, so that such content can be removed.
	  ○ TikTok uses text-based technologies, including keyword lists and natural language processing, to detect language used to promote extremist ideologies or hate groups, which enables the company to find near or exact matches of terms such as slurs (or even emoji combinations) and to remove them from comments, video captions, and profile descriptions.
	  ○ Where TikTok has previously detected content that violates its policies on hate or extremism, TikTok uses de-duplication and hashing technologies that enable it to recognise copies or near copies of such content. TikTok works with external groups,

	such as Tech Against Terrorism, to help more quickly detect and remove hate or violent extremist content that has already been identified off-platform. • 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 Trust and Safety 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. • User reports: 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 US 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). TikTok also has regional advisory councils beyond the US, including the Middle East, EU, Brazil and Asia Pacific[1] In 2022, TikTok announced a partnership with Tech Against Terrorism to prevent violent extremism. It also recalled that it may consider off-platform behaviour to identify violent extremist organisations and individuals and take action on their accounts (Bailliencourt, 2022).

[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

	TikTok is not a member of the GIFCT (although it has applied for membership).
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	TikTok's system counts the number of times an account has violated the Community Guidelines, and for each of those violations, the user receives a strike. Strikes are counted by policy area or by feature (for example, comments or direct messages). Accounts will receive a strike based on the severity of the policy violation. TikTok counts the strikes until an account reaches the threshold for a permanent account ban. For repeated violations or depending on the severity of a single violation, TikTok may permanently ban the account. In some cases, for violations when using certain features such as LIVE or direct messages, TikTok may temporarily restrict access to the feature while the content is under review to ensure that a user does not immediately re-engage in violative behaviour. When TikTok finds an account that belongs to a violent and extremist organisation or individual, it closes it immediately. TikTok may consider off-platform behaviour to identify violent extremist organisations and individuals and take action on their accounts. In limited emergency situations, TikTok will disclose user information without legal process when it has reason to believe, in good faith, that the disclosure of information is necessary to prevent the imminent risk of death or serious physical injury to any person.
7. Does the service issue transparency reports (TRs) on TVEC?	Yes (TikTok, 2023). TikTok's Community Guidelines Enforcement Report is updated quarterly with data about all of the platform's Community Guidelines content policy categories, including on TVEC. TikTok's latest transparency report, as well as an archive of earlier reports, are available at: https://www.tiktok.com/transparency/en/community-guidelines-enforcement In addition to that, TikTok published this year its first transparency report under the EU Regulation 2021/784 addressing the dissemination of terrorist content online (TCO), covering the period from 1 June to 31 December 2022. That report is available at: https://www.tiktok.com/transparency/en/tco-report/ TikTok also published its first Digital Services Act Transparency Report for the European Economic Area in

	2023. That report is available at: https://www.tiktok.com/transparency/en/dsa-transparency/

8. What information/fields of data are included in the TRs?	TikTok's last transparency report, which covers Q1 2023, includes the following metrics:
	• Removal rate of videos (%)
	• Number of videos removed (total v. by automation)
	• Number of videos restored
	• Removal rates (proactive removal rate, removal rate within 24 hours, removal rate before any views) broken down by policy violation (including violent extremism, violent and graphic content, hateful behaviour)
	• Percentage of video removed and removal rates, broken down by sub-policy:
	  ○ Violent extremism: break down between 'violent extremist organisations and individuals' and 'threats and incitement to violence'
	  ○ Hateful behaviour: breakdown between 'hateful ideology' and 'attacks and slurs on the basis of protected attributes'
	  ○ Violent and graphic content: no breakdown provided
	• Number of videos removed broken down by market/country
	• Human moderation language distribution (% of moderators assigned)
	• Response time to community-reported content (less than 2 hours, 2 to 8 hours, 8 to 24 hours, more than 24 hours)
	• Number of accounts removed (fake accounts removed, other accounts removed, accounts suspected to be under the age of 13 removed)
	• Information on fake engagement (fake likes removed, fake followers removed, fake follow requests prevented, fake likes prevented)

	• Spam account activity • Number of ads removed (ads removed due to account actions v. ad removals) On violent extremism in particular, TikTok informs that in Q1 2023, of all videos removed, 1.4% violated this policy, which is stable compared to Q4 2022. Of these videos, 94.9% were removed before they were reported, 77.4% were removed before any views, and 85.9% were removed within 24 hours of being posted.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	No information provided.
10. Frequency/timing with which TRs are issued	TikTok's Community Guidelines Enforcement Report is published quarterly. TikTok's EU Terrorist Content Online Regulation Transparency Report is published annually (as required by that law).
11. Has this service been used to post TVEC?	Yes, see Section 8 above. TikTok has been used to spread videos glorifying the Christchurch terrorist attack, and supporting the actions of racially motivated mass shooters (O'Connor & Smith, It is (still) shockingly easy to find terrorist content on TikTok, 2023).

12. Main changes since last Report	• TikTok updated its Community Guidelines in April 2023 and now uses different categories of violating content:
	  ○ The former category 'Violent extremist organisations and individuals' was replaced by a new category titled 'Violent and hateful organisations and individuals'. In this new policy, terrorist organisations are not specifically included anymore, whereas before they were clearly defined and prohibited. Now, the policy also defines 'violent political organisations' and 'hateful organisations'. In addition, the definition of 'violent extremists' references the United Nations' designation list.
	  ○ The category 'Threats and incitement to violence' was changed to 'Violent behaviours and criminal activity'.
	  ○ The category 'Hate speech', 'Slurs' and 'Hateful ideology' were regrouped together into 'Hate speech and hateful behaviours'.

	○ All categories now include examples of content and behaviour that are allowed. • TikTok's last transparency report on 'Community Guidelines Enforcement' (covering Q1 2023) features additional metrics: ○ Percentage of video removed and removal rates, broken down by sub-policy: ▪ Violent extremism: break down between 'violent extremist organisations and individuals' and 'threats and incitement to violence' ▪ Hateful behaviour: breakdown between 'hateful ideology' and 'attacks and slurs on the basis of protected attributes' ▪ Violent and graphic content: no breakdown provided ○ Human moderation language distribution (% of moderators assigned) ○ Response time to community-reported content (less than 2 hours, 2 to 8 hours, 8 to 24 hours, more than 24 hours) • TikTok issued its first transparency report under the EU Regulation 2021/794 addressing the dissemination of terrorist content online (TCO), covering Q2 2022. • In early 2023, TikTok also announced an update on efforts to enhance transparency with the research community, currently available to eligible academic researchers in the US and EU. An overview of TikTok's Research API is available at https://www.tiktok.com/transparency/en-us/research-api/ and more detailed information, including how to apply, can be found on the TikTok for Developers website at https://developers.tiktok.com/products/research-api/

8. 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) 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.
12. Main changes since last Report	No main changes since last Report.

9. 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 Acceptable Use Guidelines, 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: o 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 o 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: o engaging in or promoting acts on behalf of a terrorist organisation or violent extremist group; o recruiting for a terrorist organisation or violent extremist group; o providing or distributing services (e.g., financial, media/propaganda) to further a terrorist

	organisation's or violent extremist group's stated goals; o 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: o threatening to kill someone; o threatening to sexually assault someone; o threatening to seriously hurt someone and/or commit a violent act that could lead to someone's death or serious physical injury; o 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 Acceptable Use Guidelines are available at https://explore.zoom.us/en/community-standards/, and ToS at https://explore.zoom.us/en/terms/#:~:text=Zoom%20will%20only%20access%2C%20process,%3B%20(iv)%20as%20required%20by
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 Acceptable Use Guidelines. Zoom may delete any customer content, at any time without notice to the user if it becomes aware that it violates any provision of its ToS or any applicable laws
4.1 Notifications of removals or other enforcement decisions	Zoom notifies account owners when actions are taken on their account. Violations of the Acceptable Use Guidelines that result 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, hash-sharing/URL sharing database)	Zoom relies on reports to learn of alleged violations of its Acceptable Use Guidelines and ToS. When a user makes a report, 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 panellists serve for one-year terms and come from a diversity of backgrounds, experience levels, tenures, and departments at Zoom (Zoom, 2023). 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). Zoom is a member of the GIFCT.
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. In some circumstances (for example, credible threats of violence), Zoom might block a user on the first report of a violation. In other instances, it might record a strike, and block a user only when they have accumulated a certain number of strikes. 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, 2023). Zoom issues monthly Acceptable Use Guidelines Enforcement reports. The latest information available concerns July 2023.
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 / Zoom Events host suspended, Event suspended, duplicate or dismissed.
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. Actions taken by Zoom are defined as follows: • 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/ZoomEvents Host(s) Suspended: Zoom blocked one or more hosts of OnZoom or ZoomEvents. • 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.
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. During the COVID-10 pandemic, several 'Zoombombing' incidents have been reported, where individuals would intrude into Zoom videoconferences and share extremist and violent content to disrupt the meeting (Anti-Defamation League, 2020).
12. Main changes since last Report	No main changes since last Report.

10. Snapchat

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	Snapchat's Community Guidelines (Snapchat, 2023), under the section titled 'Hateful Content, Terrorism, and Violent Extremism', Snap states that: • Terrorist organisations, violent extremists, and hate groups are prohibited from using its platform. There is no tolerance for content that advocates or advances terrorism or violent extremism. • 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. In 2023, Snapchat published an Explainer (Snapchat, 2023) with additional information on its policy regarding 'Hateful Content, Terrorism, and Violent Extremism'. It defines TVEC as follows: "Snapchat's prohibitions against Terrorism and Violent Extremism extend to all content that promotes terrorism or other violent, criminal acts committed by individuals or groups to further ideological goals. These rules also prohibit any content that promotes or supports foreign terrorist organisations or extremist hate groups – as designated by credible, third-party experts – as well as recruitment for such organisations or violent extremist activities." The Explainer also states that hateful content and activities that support terrorism or violent extremism have no place on Snapchat. Snapchat's policies operate to create an environment that supports and prioritises the safety of Snapchatters, and to protect communities from violence and discrimination. It is never acceptable to engage in hateful conduct, including the use of hate speech or hate symbols, which means any imagery that is intended to represent hatred or discrimination towards others (including those featured in the hate symbols database maintained by the Anti-Defamation League (Anti-Defamation League, 2023)). Prohibited activities under this policy may be reported to law enforcement. Moreover, in March 2023, Snapchat published 'Content Guidelines for Recommendation Eligibility', to inform content creators how to be eligible for algorithmic recommendation. Under these Guidelines 'Hateful Content, Terrorism, and Violent Extremist Content' is strictly prohibited. The following type of content is considered 'sensitive' and may be eligible for recommendation but may be blocked for certain users (based on

	their age, location or personal preferences): hate speech, hate symbols, terrorism or violent extremism in the contexts of news, counter speech, education or respectful public discourse, as long as slurs and hate symbols are obscured. Snapchat states that content is not eligible for recommendation to anyone if it is harmful, shocking, exaggerated, deceptive, intended to disgust, or in poor taste. Accounts that repeatedly or egregiously violate the eligibility criteria may be temporarily or permanently disqualified from recommendations. Accounts that repeatedly or egregiously violate our Community Guidelines or Terms of Service may be temporarily or permanently suspended (Snapchat, 2023). Lastly, under the section titled 'Threats, Violence & Harm', Snapchat's Community Guidelines state 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, https://www.snap.com/en-GB/community-guidelines, and https://snap.com/en-GB/content-recommendation-guidelines#introduction.
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	Not applicable. Snapchat's live feature is only available to official partners in the context of specific events.
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?	Snapchat's Terms of Service state 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. The Explainer on 'Hateful Content, Terrorism, and Violent Extremism' adds that users engaged in terrorist activities or violent extremism will lose account privileges. In addition, certain information related to violations of these policies may be referred to law enforcement. Snap supports the Santa Clara Principles on Transparency and Accountability in Content Moderation (Santa Clara University's High Tech Law Institute, 2021), 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, however, it is not clear whether this is practically applied or not.

4.1 Notifications of removals or other enforcement decisions	Snapchat has a notice process to ensure that Snapchatters have a clear understanding of why an action has been taken against their account. The reporting party and the reported party are notified separately about the enforcement action (Snapchat, 2023). Snapchat also notifies users when it receives legal process seeking their account information, with exceptions for cases where it is legally prohibited from doing so, or when it believes that exceptional circumstances exist (like child sexual exploitation or an imminent risk of death or serious bodily injury) (Snapchat, 2022).
4.2 Appeal processes against removals or other enforcement decisions	Users can appeal enforcement decisions regarding content and accounts. The reporting party and the reported party are notified separately about how to submit an appeal. If the appeal is successful, Snapchat will notify the user via email and the reported content will be made eligible for recommendation, or the account will be reinstated. If the appeal is denied, Snapchat will notify the user via email and the content will not be eligible for recommendation, or the account will be deleted (Snapchat, 2023).
5. Means of identifying TVEC (for example, monitoring algorithms, user-generated, human (staff) reviewers, hash-sharing/URL sharing database)	Users can report accounts and content for potential violations of Snapchat's Community Guidelines, via submitting a confidential report directly to Snapchat's Trust & Safety team, who work on a 24/7 basis and are trained to evaluate the report; take appropriate action according to its policies; and notify the reporting party of the outcome – typically within a matter of hours. Also, Snapchat explains that its teams consult the expertise and work of civil rights organisations, human rights experts, law enforcement agencies, NGOs, and safety advocates. In addition to in-app reporting, Snapchat also offers online reporting options through its support site. Snapchat uses a combination of automated tools and human review to moderate our public content surfaces (such as Spotlight, Public Stories, and Maps) – including machine learning tools and dedicated teams of real people – to review potentially inappropriate content in public posts. Since 2022, Snapchat has been investing heavily in AI and machine-learning tools to detect violating content. On Spotlight, for example, where creators can submit creative and entertaining videos to share with the broader Snapchat community, all content is first reviewed automatically by artificial intelligence before gaining any distribution. Once a piece of content gains more viewership, it's then reviewed by human

	moderators before it is given the opportunity to reach a large audience. This layered approach to moderating content on Spotlight reduces the risk of spreading misinformation, hate speech, or other potentially harmful content, in addition to promoting a fun, positive, and safe experience for everyone. Additionally, Snapchat uses proactive harm-detection technology on public and high-visibility surfaces – such as Stories – to help identify harmful content, and we use keyword filtering to help prevent harmful content (such as accounts trying to advertise illicit drugs or other illegal content) from returning in search results. In 2022, Snapchat rebuilt its Safety Advisory Board, expanding membership to include a diversity of geographies, safety-related disciplines and areas of expertise. The Board was created in 2018 to independently educate, challenge, raise issues to, and advise Snapchat on how to help keep the Snapchat community safe (Snapchat, 2022). Snapchat is not a member of the GIFCT.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Snapchat considers a combination of factors to determine the appropriate penalties for violations of the Community Guidelines. The most important of these factors are the severity of the harm and any relevant history by the Snapchatter of previous violations. It applies a risk-based approach to distinguish the most severe harms from other types of violations that may not rise to the same level of seriousness. Accounts that are used primarily to violate the Community Guidelines or to perpetrate serious harms will immediately be disabled. Examples include accounts engaged in serious bullying or harassment, impersonation, fraud, promotion of extremist or terrorist activity, or otherwise to engage in illegal activity. For all other violations of our Community Guidelines, Snap applies a three-part enforcement process: • Step one: the violating content is removed. • Step two: the Snapchatter receives a notification, indicating that they have violated our Community Guidelines, that their content has been removed, and that repeated violations will result in additional enforcement actions, including their account being disabled. • Step three: our team records a strike against the Snapchatter's account. A strike creates a record of violations by a particular Snapchatter. Every strike is accompanied by a notice to the Snapchatter; if a

	Snapchatter accrues too many strikes over a defined period of time, their account will be disabled (Snapchat, 2023). In particular, when hateful content is reported, Snapchat will remove any violating content and users who engage in repeated or egregious violations will have their account access locked. As an additional measure, it encourages Snapchatters to block any users who make them feel unsafe or uncomfortable. Users engaged in terrorist activities or violent extremism will lose account privileges. In addition, certain information related to violations of these policies may be referred to law enforcement (Snapchat, 2023). Some 'sensitive' content, while still be eligible for algorithmic recommendation, may be blocked for certain users based on their age, location, preferences, or other criteria. In the category of 'Hateful Content, Terrorism and Violent Extremism', sensitive content includes hate speech, hate symbols, terrorism or violent extremism in the contexts of news, counter speech, education or respectful public discourse, as long as slurs and hate symbols are obscured (Snapchat, 2023).
7. Does the service issue transparency reports (TRs) on TVEC?	Yes. Snapchat's last transparency report (Snapchat, 2023) covers the period from July to December 2022. In a new section, titled 'Analysis of Content and Account Violations', Snapchat assesses major data changes relative to our previous reporting period. Lastly, to improve consistency, the ordering of the report now mirrors Snapchat's Community Guidelines, as suggested by Snapchat's Safety Advisory Board.
8. What information/fields of data are included in the TRs?	Snapchat's transparency report includes: • Overall violative view rate (VVR) • Total number of content and account reports, broken down by total content enforced and total unique account enforced • Total number of content and account reports, broken down by category of violating content (including threats and violence) • Number and percentage of content enforced, broken down by category of violating content • Number of unique accounts enforced, broken down by category of violating content

	• Turnaround time, broken down by category of violating content • Number of account removals for violations of Snapchat's prohibition of terrorist and violent extremist content • Country-by-country information, including the number of account deletions for 'Terrorism'. Snapchat indicates that 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 (Snapchat, 2021).
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	• Violative View Rate: The proportion of all Snaps (or views) that contained content that violated Snapchat's Community Guidelines during the reporting period. During the last reporting period, the VVR was 0.03 percent, which means that out of every 10,000 views of content on Snapchat, three contained content that violated its guidelines. • Content and account reports: Total number of content pieces reported and accounts reported to Snapchat. • Enforcement (enforced): An action taken against a piece of content or account (e.g. deletion, warning, locking, preservation). • Total content enforced: The total number of pieces of content (e.g., Snaps, Stories) that were enforced against on Snapchat. • Total unique accounts enforced: The total number of unique accounts that were enforced against on Snapchat. For example, if a single account was enforced against multiple times for various reasons (warned for posting false information, and then later deleted for harassing another user), only one account would be calculated in this metric. Both enforcement actions would, however, be included in the "Overview of Content and Account Violations" table, with one unique account enforcement for "False Information" and one unique account enforcement for "Harassment and Bullying." • Turnaround time: The duration of time between when our Trust & Safety teams first start to review a report (usually when a report is submitted) to the last enforcement action timestamp. If multiple rounds of review occur, the final time is clocked at the last action taken.

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.

12. Main changes since last Report	• In 2023, Snapchat updated its policies on TVEC:
	  ○ Snapchat published an Explainer on its 'Hateful Content, Terrorism, and Violent Extremism' policies. It notably provides a definition for TVEC, as well as more details about TVEC-specific detection means and enforcement methods.
	  ○ Snapchat published 'Content Guidelines for Recommendation Eligibility', to inform content creators how to be eligible for algorithmic recommendation. Under these Guidelines 'Hateful Content, Terrorism, and Violent Extremist Content' is strictly prohibited. The following type of content is considered 'sensitive' and may be eligible for recommendation but may be blocked for certain users (based on their age, location or personal preferences): hate speech, hate symbols, terrorism or violent extremism in the contexts of news, counter speech, education or respectful public discourse, as long as slurs and hate symbols are obscured.
	  ○ Snapchat updated its Community Guidelines which now include a category of violating content titled 'Hateful content, Terrorism, and Violent Extremism' (formerly 'Terrorism, Hate Groups and Hate Speech'), which puts an emphasis on violent extremism.
	• Snapchat now provides detailed information about its detection and enforcement methods:
	  ○ Snapchat has been investing heavily in AI and machine-learning tools to proactively detect violating content. It uses a layered approach combing automated tools and human review.
	  ○ Snapchat takes into account the severity of the harm and any relevant history by the user of previous violations. It has established a strike system that creates a record of violations for users.
	  ○ Accounts that are used primarily to violate the Community Guidelines or to perpetrate serious harms, such as promoting terrorist and extremist activity, are immediately be disabled.

	• Snapchat has now established notice and appeal processes, and provides detailed information on how they work. • Snapchat's transparency report features an updated glossary which provides detailed information on the terminology used and the methodology used for calculating metrics.

11. Douyin

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, Douyin's User Service Agreement (Douyin, 2023) prohibits users from producing, copying, publishing, or disseminating the following: • Promoting terrorism, extremism, Islamism, or inciting the implementation of terrorist and extremist activities; • Promoting ethnic hatred, ethnic discrimination, and undermining ethnic unity; • Spreading and disseminating violence, murder, terror, or instigating crimes; • Showing blood, horror, cruelty etc. that cause physical and mental discomfort; • Inciting discrimination; • Intimidating or threatening others with violence, and committing cyber violence; • Any content of violence and/or self-harm; • Any content that threatens life and health and uses dangerous equipment; • Encouraging or inducing others to participate in dangerous activities that may cause personal injury or death.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.douyin.com/draft/douyin_agreement/douyin_agreement_user.html?id=6773906068725565448
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No. Douyin's User Service Agreement apply to all content available on Douyin, including live broadcasts.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there	See section 6 below.

notifications of removals and appeal processes against removal decisions?	
4.1 Notifications of removals	No information is specified.
4.2 Appeal processes against removal decisions	If a user does not agree with an enforcement action taken by Douyin, he or she can submit an appeal, through the designated entrance in the Douyin App, or by email to jubao@douyin.com.
5. Means of identifying TVEC (for example, monitoring algorithms, user-generated, human (staff) reviewers, hash-sharing/URL sharing database)	Douyin uses a combination of automated tools and human reviewers to moderate content on its platform. In 2020, Douyin reportedly had over 20,000 content moderators (Meihan, China's Content Moderators Are Overworked and Chronically Stressed, 2022). 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 L. , 2019). Douyin is not a member of the GIFCT.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	In case of violation of the User Service Agreement, Douyin states that it has the right to take advance warning, refuse to publish, immediately stop transmitting information, delete content, ban the publication of content or comments for a short period, and restrict part or all of an account as appropriate, permanently close an account, and take any other measures in accordance with laws and 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?	Yes. In 2021, Douyin was fined for the dissemination of pornographic and vulgar content by the Chinese government. Such content including live-streaming games involving violence and terror (The Straits Times, 2021).

12. Main changes since last Report	Douyin was not included in previous Reports.

12. 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, bots, sticker sets, etc. 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?	Telegram states that if they receive a court order that confirms a user is a 'terror 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.). When it does, Telegram states that it will include it in a semi-annual transparency report published at: https://t.me/transparency.
4.1 Notifications of removals or other enforcement decisions	No notifications are specified.
4.2 Appeal processes against removals or other enforcement decisions	Users can send an appeal using @Spambot. Although this appeal process seems to be primarily reserved for cases of phishing and spam, the policy also encompasses "other kinds of abuse and violations of Telegram's Terms of Service". Telegram has designated a third party, European Digital Services Representative (EDSR), for TCO-related communications. Users whose publications were taken down in connection with the TCO Regulation can request details on why their publications were considered terrorist and how to challenge the removal by contacting EDSR or the @EURegulation bot on Telegram.
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 either by contacting abuse@telegram.org or by using the 'report' button inside the app. 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. For instance, in August 2023, Telegram banned 11,431 terrorist bots and channels. Telegram may also use automated algorithms to analyse messages in cloud chats, however the ToS specifically mention these for stopping spam and phishing. It is not clear weather automated tools are also used for identifying TVEC. Since the entry into force of the EU Regulation 2021/784 addressing the dissemination of terrorist content online (TCO) in 2022, authorities in EU countries can send removal requests for terrorist content, if it is discovered on Telegram's public platform. Telegram has designated a third party, European Digital Services Representative (EDSR), to assist with TCO-related communications. Telegram is not a member of the GIFCT.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Telegram's ToS, in Section 5.3 titled 'Spam and Abuse', state the following: "To prevent phishing, spam and other kinds of abuse and violations of Telegram's Terms of Service, our moderators may check messages that were reported to them by their recipients. If a spam report on a message you sent is confirmed by our moderators, your account may be limited from contacting strangers – temporarily or permanently. You can send an appeal using @Spambot. In case of more serious violations, your account may be banned." It is not clear what is meant by "more serious violations".
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. Telegram removed 8747 ISIS channels and bots during the month of June 2023). Telegram states that, if it receives a court order that confirms a user is a 'terror suspect', and disclose that user's IP address and phone number to the relevant authorities, it will include it in a semi-annual transparency report published at: https://t.me/transparency
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. In the November 2015 attacks in Paris, France, a cell of IS affiliates used the application to communicate internally, killing 130 people in the deadliest jihadist attack in French history. Among the most tragic ones, there is also the December 2016 truck-ramming attack on a Christmas market in Berlin, Germany, and the mass shooting at the Reina Nightclub in Istanbul, Türkiye. In each case, the attackers are believed to have received instructions via Telegram's secret chat function (Bennett, 2019). Telegram is also used by far-right extremists to advocate terrorism and glorify terrorist acts (Hayden, 2019).
12. Main changes since last Report	• Telegram states that, if it receives a court order that confirms a user is a 'terror suspect', and disclose that user's IP address and phone number to the relevant authorities, it will include it in a semi-annual transparency report published at: https://t.me/transparency • Telegram has designated a third party, European Digital Services Representative (EDSR), to assist with TCO-related communications.

13. Kuaishou/Kwai

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	Kuahishou/Kwai's Community Guidelines (Kwai, 2022), last updated in February 2022, define 'terrorist organisations' as 'any non-state actor that engages in, advocates, or lends substantial support to acts of violence that could cause death, injury, or serious harm to civilians with the intent to coerce, intimidate, or influence a civilian population, government, or international organisation to achieve a political, religious, or ideological aim'. Under Section 7 titled 'Dangerous Individuals and Organisations', Kwai states that terrorism and organised crime present a threat to the safety of users. Therefore, it does not allow dangerous individuals and organisations to use its services to promote terrorism, crime or other harmful

	activities, and once identified, it may suspend or terminate such accounts and notify the relevant legal authorities. Under this policy, users may not, nor may they permit any other person to upload, share, publish, transmit, stream or otherwise make available on or through the service any content that: • Engages in terrorist activity or praises, glorifies, supports, or promotes terrorist organisations (as defined above); • Engages in or promotes organised criminal activity which includes groups dedicated to committing crimes or cause other types of harm with the use of violence. These crimes include but are not limited to: homicide, human trafficking, organ trafficking, drug trafficking, arms trafficking, criminal impersonation, money laundering, extortion, kidnapping, and sexual exploitation; • Contains names, symbols, logos, flags, slogans, uniforms, gestures, portraits, or other objects meant to represent terrorist, organised crime groups and their members with the intent or promote those groups. Under Section 2 titled 'Violent and Graphic Content', Kwai does not allow content that is extremely graphic, or normalises or glorifies violence, since this type of content may also lead to real-world violence. If there is any threat to the users' safety, Kwai may ban the account and report the threat to the appropriate legal authorities. Users may not, nor may they permit any other person to upload, share, publish, transmit, stream or otherwise make available through the service any content that: • Depicts violent deaths or accidents; • Depicts or promotes violent fights; • Depicts corpses (except in educational, documentary, scientific or artistic context); • Depicts or promotes extremely violent, bloody or graphic content, such as torture, dismemberment, visible innards or severe injuries; • Depicts graphic content that may shock, cause nausea or discomfort, such as human excretion, gory jump scare and uncensored childbirth; • Depicts animal slaughter; • Depicts or promotes animal abuse. Under Section 6 titled 'Hateful Behaviour', Kwai states prohibits discrimination based on protected attributes (age,

	caste, gender or gender identity, immigration status, nationality, race or ethnicity, religion, serious disease or disability, and sexual orientation). Users may not, nor may they permit any other person to upload , share, publish, transmit, stream or otherwise make available through the service any content that: • Incites hatred against any person based on their protected attributes; • Does or intends to attack, threaten, incite violence against, or dehumanise an individual or group based on protected attributes; • Could be interpreted as being hateful, harassing, abusive, racially or ethnically offensive, discriminatory towards a particular class, individual or group, defamatory, humiliating, or disrespectful to other people; • Contains degrading terms intended to insult an individual or group based on a protected attribute ('slurs'); • Promotes hateful ideologies, i.e., ideologies that are substantially based on shared hostility against people of one or more protected attributes, which can have a negative impact on the platform's openness and tolerance and may lead to real-world violence against groups; • Praises or belittles violent tragedies (events that led to several deaths and serious injuries) and their victims. Finally, Kwai developed an 'Election Policy' to specifically address harmful content in the context of an electoral process. Under this policy, Kwai will label or take down content that aims to spread and/or generate an atmosphere of violence in the electoral process, thus making refrain from exercising their right to vote. This category also encompasses content promotes and incites violent actions to obstruct electoral activities. More generally, Kwai'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 (Kwai, 2022).
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.kwai.com/safety?id=community; and https://app.kwai.com/agreement/service-terms

3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	Neither Kwai's Terms of Service nor Community Guidelines contain specific provisions for livestreamed content. However, the Help Section of Kuaishou's website (Kuaishou, 2023), the version of the app available in China, provides information on how to activate the live broadcast function. Users have to apply to get live broadcast permission and upload a personal ID card in accordance with national requirements. If the account making the request has recently committed a violation, or too many historical violations, then the request may be refused. Minors under the age of 18 cannot activate the live broadcast function. In a live room, super administrators and administrators can set sensitive words to facilitate content moderation. They can also block or kick out viewers. These changes have occurred in a context of increasingly more strict requirements for livestream content in China and increased scrutiny from the Chinese authorities (Tindall, What are the current regulations for live streaming in China, 2022).
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?	In case of violation of any provisions of the Community Guidelines, Kwai may suspend or terminate a user's access to part or whole of the service, restrict access to the user's content, remove the content, or take any other appropriate action in accordance with the ToS without notice. If necessary, Kwai may cooperate with law enforcement agencies to investigate any illegal acts on the service. Kwai also 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	When a video is removed, the user receives a notification specifying the reason for removal.
4.2 Appeal processes against removals or other enforcement decisions	Kwai has an appeal process in place. Users have the possibility to appeal a video removal decision – the access to the appeal is included in the notification. After an appeal is submitted, Kwai's moderation team will review it and take action accordingly. If there is no violation, the video will be reinstated.
5. Means of identifying TVEC (for example, monitoring algorithms, user-	Kwai has a user report mechanism that allows users to report violations (report@kwai.com). In particular, users can report another user, a video, a comment, or a private message, and when doing so they should provide a report reason. These

generated, human (staff) reviewers, hash-sharing/URL sharing database)	reports are verified and processed by staff moderators. In December 2022, Kwai updated its report mechanism by providing new report portals, refining report reasons, and opening windows for user feedback. Users can now report an inappropriate video by long pressing it, in addition to the existing 'sharing' icon. For account report, users can click the 'sharing' icon in addition to the existing '3-dot' icon in the user's profile page. In addition, users can now specify the violative scenario (video, comment, live, etc.) and give their reporting experience a satisfaction rate. In addition, Kwai indicates that it uses machine learning technology in order to review content on a mass scale, including videos, comments, livestreams and advertising, to detect any violation. When content is uploaded on the platform, it is checked by machine learning algorithms. If a violation is suspected, is then reviewed by a team of local moderators who are fluent in the language and contextually aware. 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 L. , 2019). Kwai is not a member of the GIFCT.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	In case of violation of any provisions of the Community Guidelines, Kwai may suspend or terminate a user's access to part or whole of the service, restrict access to the user's content, remove the content, or take any other appropriate action in accordance with the ToS without notice. If necessary, Kwai may cooperate with law enforcement agencies to investigate any illegal acts on the service.
7. Does the service issue transparency reports (TRs) on TVEC?	No. Kuaishou International Business (KSIB) has been publishing semi-annual transparency reports since H1 2021. The transparency reports account for both Kwai and Snack Video, its two platforms tailored for users outside mainland China, but not for Kuaishou.
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	Takedown within 24 hours refers to removing violative content within 24 hours of its posting.

	Proactive detection refers to identifying and removing violative content before users report it. Takedown with zero view refers to removing violative content before anyone has viewed it.
10. Frequency/timing with which TRs are issued	On a semi-annual basis.
11. Has this service been used to post TVEC?	Unknown.
12. Main changes since last Report	• Kuaishou International Business (KSIB), Kwai's parent company, released its first transparency report in H2 2021, covering both Kwai and Snack Video, its two short video platforms tailored for users outside of mainland China. • Kwai updated its Community Guidelines that now outline precise categories of violating content, including 'Dangerous Individuals and Organisations'. They provide a definition of 'terrorist organisations' and specific examples. • Kwai also refined its user report mechanism. • Kuaishou, the Chinese version of the app, has implemented stricter requirements for livestreamed content, as the Chinese authorities have issued new guidelines for the live-streaming industry.

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[52] 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://www.qq.com/contract.shtml[53]
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:	No procedure is specified.

are 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)	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 L. , 2019). QQ International is not a member of the GIFCT.
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.
12. Main changes since last Report	No main changes since last Report.

14. 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 false, fake, harassing, defamatory, offensive, abusive, threatening, racially discriminatory, slanderous, leaking privacy, pornographic, obscene, malicious, plagiarism, violent, gore, suicide, self-mutilation or otherwise illegal. Under Article 4.2 of Weibo's ToS, users can edit the account name, avatar, profile, etc. but they must not contain illegal or harmful information. And under Article 4.9, when using Weibo, users shall not violate the laws and regulations of the People's Republic of China and relevant international treaties or rules.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.weibo.com/signup/v5/protocol; and https://m.weibo.cn/c/regagreement?from=h5&lang=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?	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 messages, participation in topics, participation in activities, marketing information, reporting and complaints, etc.). Under Article 4.11, if Weibo discovers or receives third-party report or complaint that the user violates Article 4 (entitled 'Rules of Use'), it has the right to request the user to make corrections within a time limit; or take all necessary measures without notice to reduce or eliminate the impact of the user's misconduct, and will notify the user after processing as much as possible. Such measures include but are not limited to changing, blocking or deleting relevant content, warning the violating account, restricting or prohibiting some or all functions of the violating account, suspending, terminating, cancelling the user's right to use Weibo. Lastly, in 2022, Weibo announced that it would start publishing users' province or municipality (for users in China) and IP

	locations (for users overseas) on their account pages and when they post comments. In a notice, Weibo said that the settings are designed to "reduce bad behaviour such as impersonating parties involved in hot topic issues, malicious disinformation and traffic scraping, and to ensure the authenticity and transparency of the content disseminated" (Reuters, 2022).
4.1 Notifications of removals or other enforcement decisions	Weibo's ToS mention a notification system that remains at Weibo's discretion for content removal. Weibo will notify the user after processing as much as possible.
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)	Upon Article 4.12 of its ToS, Weibo has a reporting mechanism that allow users to report illegal or infringing content uploaded by other users. 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 L. , 2019). Weibo is not a member of the GIFCT.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Under Article 4.3 of Weibo's ToS, if the user violates Article 4, Weibo has the right to take measures such as not registering the user, notifying a deadline for correction, cancelling the registered account, suspending or terminating the 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?	Yes. The Christchurch shooting was posted on Weibo (Kenny, 2019).
12. Main changes since last Report	• Weibo has a notification system to inform users about enforcement actions, but it remains at its sole discretion. • In 2022, Weibo announced that it would start publishing users' province or municipality (for users in China) and IP locations (for users overseas) on their account pages and when they post comments, with a view to "reduce bad behaviour".

15. 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 or property; • 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://international.qq.com/privacy/mobile/terms/terms.html?language=English, https://www.tencent.com/en-us/zc/termsofservice.shtml, https://www.tencent.com/en-us/zc/acceptableusepolicy.shtml[54]
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or	QQ broadly states that it may review (but make no commitment to review) content (including any content posted by users) or third party services made available through QQ to determine whether or not they comply with QQ's policies, applicable laws and regulations or are otherwise objectionable. QQ may remove or refuse to make available or link to certain content or third-party services if they infringe intellectual property rights, are obscene, defamatory or abusive,

other enforcement decisions and appeal processes against them?	violate any rights or pose any risk to the security or performance of its services. Also, QQ reserves the right to block or remove content for any reason, as required by applicable laws and regulations.
4.1 Notifications of removals or other enforcement decisions	No notifications are specified.
4.2 Appeal processes against removals or other enforcement decisions	No appeal processes are specified.
5. Means of identifying TVEC (for example, monitoring algorithms, user-generated, human (staff) reviewers, hash-sharing/URL sharing database)	Users can report violations through the reporting function. QQ also uses tools to proactively discover policy violations. 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 L. , 2019). In particular, research shows that QQMail, a mail service integrated to the QQ application, uses automated systems to filter content by keywords (Knockel & Ruan, Measuring QQMail's automated email censorship in China, 2021). QQ is not a member of the GIFCT.
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. Main changes since last Report	No main changes since last Report.

16. 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), ethnic 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 L. , 2019). iQIYI is not a member of the GIFCT.
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 temporarily suspend or permanently terminate a user's account, and interrupt or terminate the continued provision of iQIYI products and/or services without any liability.
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. Main changes since last Report	No main changes since last Report.

17. 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, 2023), Pinterest states that its platform is not a place for violent content, groups or individuals. Pinterest limits the distribution of or removes 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 its 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?	In 2022, Pinterest launched Pinterest Studio app, a live-streaming tool for video creators (Perez, 2022). All creators are eligible to apply, but they must fulfil the following requirements: • Be a business account holder • Be located within the United States or Canada • Be 18 years or older • Be active on Pinterest:

	○ 250 or more followers ○ 150 or more saves in the last 30 days ○ 3 or more video Pins created in the last 30 days Then, in order to create an episode, users must submit their idea at least seven days before the foreseen livestream. Pinterest will review the information and decide whether to approve the episode or not. (Pinterest, 2023)
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	If a user thinks that Pinterest made an enforcement error, they have the possibility to submit an appeal via the Help Centre or by clicking the one-click appeal link in the enforcement notice email. Pinterest reviews appeal requests and updates the enforcement decision accordingly. Pinterest may limit appeals; for example, it may suspend the processing of appeals from people who frequently submit unfounded or abusive appeals, and it may limit the number of times that a particular decision can be appealed. Pinterest may also use automation to handle appeals more efficiently, for example by expanding a decision made on one Pin to other similar Pins. Appeals availability may vary for some product features or in some localities. And some Pinners may have additional appeal options or mechanisms under their local law.
5. Means of identifying TVEC (for example, monitoring algorithms, user-generated, human (staff) reviewers, hash-sharing/URL sharing database)	Pinterest has a reporting mechanism through which users can report content that violates its policies. It includes Pins, accounts, comments and photos comments, messages, and boards. When reporting, users are required to provide a reason. Pinterest uses automated tools, manual review and hybrid approaches, both to identify and take action on violating content. The automated systems may use machine learning as well as logic-based rules. Where appropriate, Pinterest may take into account information provided by trusted third parties

	(such as industry, government and security experts), and industry tools. • Automated actions: Pinterest's automated tools use a combination of signals to identify and take action against potentially violating content. For example, its machine learning models assign scores to content added to the platform. The automated tools can then use those scores to perform appropriate enforcement actions. • Manual actions: Pinterest manually acts on some Pins through its human review process. Pins actioned through this process may include those identified internally and those reported by third parties. It also includes the Pins that are reviewed and actioned by dedicated team members after a user report. • Hybrid actions: Hybrid actions include those where a team member determines that a Pin violates policy, and automated systems help expand that decision to enforce against machine-identified matching Pins. Depending on the prevalence of matching Pins, a hybrid action may result in a number of Pins actioned or none at all. Pinterest is a member of the GIFCT.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	When Pinterest becomes aware of a Community Guidelines violation, it may take the following actions: • Deactivation: When content is deactivated - such as Pins, boards, comments or user accounts - that content is no longer available to anyone on the platform. Deactivation can also be referred to as "removal;" or • Limiting distribution: When Pinterest limits the distribution of a Pin, it will continue to be accessible on Pinterest, but it won't be featured in recommendation or discovery surfaces, such as search results or the home feed. Pinterest determines whether content should be removed or limited in distribution based on how much risk of harm it poses, particularly the severity of its impact and the vulnerability of its target. Depending on the context, Pinterest may allow content that would typically be deactivated but limit its distribution so that people do not come across it accidentally. For example, it may limit distribution of content where the context is acceptable (such as condemnation or education). Pinterest may limit or remove boards or accounts for repeated violations or when they are dedicated to a policy violation. It

	may also remove an account after a single instance of a severe policy violation or if the account has repeatedly posted illegal content. If Pinterest limits or removes a board or account, that action applies to all of the Pins contained on the board or account. In addition, boards whose distribution has been limited will not be visible when viewing someone else's profile. If Pinterest deactivates an entire account, all of the content (Pins and boards) on that account also are deactivated and no longer available to anyone on Pinterest. In certain circumstances, before an account is deactivated, Pinterest may place additional restrictions on its use, such as limiting the account's ability to post or save content. Content that, according to Pinterest's systems, may not be "inspirational, relevant and safe" may be shown less often or less prominently, even if Pinterest hasn't determined that the content necessarily goes against the Community Guidelines. Pinterest may also implement additional product features to improve a Pinner's experience, such as applying sensitivity screens in situations where the systems indicate that content might not be appropriate for all audiences. Pinterest may also limit certain features on Pins, such as turning off comments or not showing related content, when they may be unsafe or when appropriate to protect minors. Finally, Pinterest may take additional enforcement measures on: • Links: Pinterest may moderate content based on links (URLs) associated with that content. For example, it may block the creation of a Pin that links to an inappropriate website, or remove or limit distribution of an existing Pin that links to an unsafe website. • Text: Pinterest may moderate content based on text associated with that content. For example, it may deactivate or limit content that contains violating text, or decline to show search results or ads in response to queries that contain policy-violating or sensitive text. (Pinterest, 2023)
7. Does the service issue transparency reports (TRs) on TVEC?	Yes. Pinterest's last transparency report covers from July to December 2022, split into two reporting periods: Q3 and Q4 2022 (Pinterest, 2023).
8. What information/fields of data are included in the TRs?	Pinterest's latest transparency reports features a separate reporting category for its 'Violent actors' policy, which encompasses extremists and terrorist organisations. Previously, TVEC was included in other broader categories.

	All of the following metrics are provided for each category of policy violation (among which are 'Violent actors'): • Reach of policy-violating Pins (content with a message): percentage of Pins seen by 0 people, <10 people, 10-100 people and > 100 people • Number of distinct images and Pins deactivations • Percentage of Pins deactivated manually v. percentage of pins deactivated with hybrid tools • Number of actioned user reports • Number of board deactivated • Number of accounts deactivated • Number of account appeals received • Number of accounts reinstated Pinterest does not provide data on deactivated Pins and boards yet but expects to include it in future reports. The report also includes: • Government information requests, broken down by country • Government content deactivation requests, broken down by country
9. Methodologies for determining/ calculating/estimating the information/data included in the TRs	• Distinct images and Pins deactivated: Organic Pins include all Pins created and saved on Pinterest that are not promoted as ads, and each of those Pins also has an image associated with it. Much of the content on Pinterest has been saved repeatedly, meaning that the same image may appear in multiple Pins. As a result, when it comes to reporting actions taken on organic Pins, Pinterest includes the number of Pins deactivated as well as the number of distinct images deactivated to provide greater insight into its moderation practices for this type of content. To avoid double-counting deactivated boards and accounts are counted separately. • Reach: To calculate this metric, Pinterest starts by looking at each policy-violating Pin deactivated in a reporting period. Then it counts the number of unique users that saw each of those Pins during the reporting period for at least 1 second before it was deactivated.

	Reach for a policy category may not add up to 100% due to rounding. • Actioned user report: 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. • Boards deactivated: When users find Pins they like or want to come back to, they can save them to boards that they have created. Over time, our users have created billions of boards. When a board is deactivated for violating policy, all the Pins on that board are also deactivated. Similarly, when Pinterest deactivates an entire account, that user's boards are also deactivated. To avoid double-counting deactivations, the count of boards deactivated does not include those from user accounts that were deactivated.
10. Frequency/timing with which TRs are issued	On a semi-annual basis. The transparency reports provide the details for each quarter.
11. Has this service been used to post TVEC?	Presumably. Pinterest reported content violating its 'Violent actors' policy, which covers extremists and terrorist organisations, for the first time in H2 2022. For example, in Q4 2022, Pinterest deactivated 4,120 distinct images and 95 accounts under this policy. However, it is impossible to tell how many, if any, specifically concern TVEC.
12. Main changes since last Report	• Pinterest provides more detailed information about its detection methods. For example, it uses machine learning models to assign scores to content added to the platform. The automated tools can then use those scores to perform appropriate enforcement actions. • Pinterest provides more detailed information about its enforcement actions. For example, it may limit or remove boards or accounts for repeated violations or after a single instance of a severe policy violation or if the account has repeatedly posted illegal content. • Pinterest's transparency reports now includes: o A separate reporting category for its 'Violent actors' policy, which encompasses extremists and terrorist organisations. Previously, TVEC was included in other broader policies. o Percentage of Pins deactivated manually v. percentage of pins deactivated with hybrid tools o Additional explanations on the methodology used to calculate the reporting metrics.

	• Pinterest introduced a live-streaming feature, only available in the United States and Canada for now, for selected creators, and subject to strict requirements. • Users can appeal any enforcement decision, and not account suspensions only.

19. Reddit

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	'Rule 1' of Reddit's Content Policy (Reddit, 2023) prohibits violent content and threats of violence, and specifically cites terrorist content as an example. Reddit defines terrorist content as "propaganda material posted by terrorists or designated terrorist organisations and their supporters, expressions of affiliation or support for terrorists or designated terrorist organisations, glorification of terrorist acts [or] content that solicits or incites a person or group to participate, commit, or contribute to terrorist activities." Rule 1 also more generally prohibits any content that encourages, glorifies, incites, or calls for violence or physical harm against an individual (including oneself), a group of people, or animals. Reddit also notes that there are sometimes reasons to post violent content (e.g., educational, newsworthy, artistic, satire, documentary, etc.) and asks its users to provide sufficient context when necessary. The following examples of prohibited violent content are provided: • Post or comment with a credible threat of violence against an individual or group of people. • Post containing mass killer manifestos or imagery of their violence. • Terrorist content, including propaganda. • Post containing imagery or text that incites, glorifies, or encourages self-harm or suicide. • Post that requests, or gives instructions on, ways to self-harm or commit suicide. • Graphic violence, image, or video without appropriate context
2. Manner in which the ToS or Community Guidelines/Standards are communicated	All of Reddit's terms and policies are available at https://www.redditinc.com/policies. Specifically, the User Agreement is available at https://www.redditinc.com/policies/user-agreement and the Content Policy is available at https://www.redditinc.com/policies/content-policy.

	The Content Policy is also sent via direct message to all users when they register a new account. 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 (called 'subreddits') that are created and moderated by users themselves, on a volunteer basis. These moderators set their own community rules, unique to each specific subreddit depending on its topic, in addition to the sitewide Content Policy. These rules are clearly marked in the sidebars of each individual community. Moderators are subject to an additional set of rules known as the Moderator Code of Conduct, available at https://www.redditinc.com/policies/moderator-code-of-conduct. (This replaced the previous Moderator Guidelines in 2022.) The Code of Conduct is sent via direct message to users when they are appointed as moderators.
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No. Reddit's Live feature, called RPAN, was discontinued in November 2022 and past streams were deleted (r/pan, 2022). Reddit no longer offers any live-streaming features and as such the previous provisions specific to livestreamed content no longer apply.
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 (or 'Admins', who are paid Reddit employees) have a variety of different methods to enforce their rules, including: • Remove the content • Send a warning to the user who posted it • Temporarily ban the user's account(s) • Permanently ban or terminate the user's account(s) • Remove privileges from, or add restrictions to, the user's account(s) The appropriate enforcement action depends on the type and severity of the violation, as well as the user's violation history. Severe violations give rise to an immediate permanent ban of the account. For other violations, a user may first receive a warning, followed by a 3-day suspension, 7-day suspension, and then a permanent ban. All Reddit Administrator enforcement actions are documented in Reddit's Transparency Report. 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), removing their posts from the community, muting them, etc.

	These actions happen independently of Reddit administrators. More information on the role and tools of Moderators can be found in the Reddit's Mod Help Centre (Reddit, 2023). Reddit's 2023 transparency report (Reddit, 2023) shows that the majority of content removals were executed within individual communities ('subreddits') by Mods. These removals are largely based on individual subreddit rules that are unique to each community and set by Mods 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. Reddit's communities as a whole are also subject to Reddit's sitewide rules. Those that repeatedly violate the rules may be subject to warnings or bans. Reddit Admins may add restrictions to Reddit communities, such as adding NSFW tags or quarantining. Additionally, volunteer user-moderators must also adhere to a supplementary set of rules known as the Moderator Code of Conduct. Moderators that fail to comply with the Code of Conduct may be subject to a range of enforcement actions, including: • Issuing warnings • Temporary or permanent suspension of accounts • Removing moderators from a community • Prohibiting a moderator from joining additional moderator teams or creating new subreddits • Removal of privileges from, or adding restrictions to, moderators' accounts • Adding restrictions to Reddit communities, such as adding NSFW tags or Quarantining • Removal of content • Banning of Reddit communities
4.1 Notifications of removals or other enforcement decisions	When Reddit removes a piece of content for violating the sitewide Content Policy or takes an associated account-level enforcement action, the account that posted the content is notified of the removal reason and provided instructions for how to appeal. The notifications aim to educate users about how and why they violated the rules, to foster and encourage positive contributions to the platform going forward. Individual content removals or account suspensions are notified via a private message. In cases where content is removed, users at large will be notified by a marker where the post or comment previously existed. In cases where otherwise non-violating content is removed in response to a

	valid legal or government request regarding local law, rather than removing the post or comment outright, Reddit may block the post from being accessible in a particular country, and notify the original poster. Such restrictions will be similarly noticed to users subject to them, noting the specific jurisdiction of the restriction. Whole subreddit removals are also tombstoned with the removal reason so that visitors may see why and when the community was banned. Moderators are additionally notified via modmail (the shared messaging system that moderators use to communicate with members of their communities) if their account or community have had restrictions placed on it for Code of Conduct violations.
4.2 Appeal processes against removals or other enforcement decisions	Users may submit an appeal within six months of receiving the notification. Reddit states that it processes appeals in a timely, non-discriminatory, diligent, and non-arbitrary manner, and reverses the original decision if it is determined that Reddit's initial assessment was incorrect. Appeal rates are documented in Reddit's Transparency Report. Additionally, restrictions applied to entire communities based on Code of Conduct violations may be lifted at a later date when the violations have been sufficiently rectified. Reddit's Moderator Code of Conduct also requires 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. Likewise, when a subreddit is quarantined, moderators may present an appeal with a detailed accounting of changes to community moderation practices (such as adding more moderators, creating new rules, employing more aggressive auto-moderation tools, adjusting community styling, etc.) The appeal should also offer evidence of sustained, consistent enforcement of these changes over a period of at least one month, demonstrating meaningful reform of the community.
5. Means of identifying TVEC (for example, monitoring algorithms, user-generated, human (staff) reviewers, hash-sharing/URL sharing database)	Reddit has a community-based, multi-tiered, democratic approach to content moderation Company employees ('Admins') are responsible for enforcing the Content Policy across the platform. Individual communities (also known as "subreddits") also have their own community-specific rules, which are created and enforced by other users based on their

	community's unique and often highly specific topics. The users who create and enforce these subreddit-specific rules are known as community moderators, or ("mods"). Moderating a Reddit community ('subreddit') is an unofficial, unpaid position. Community creators are automatically that community's first moderators, and they may appoint other users as Mods 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 Code of Conduct (Reddit Inc., 2022), which replaced the Moderator Guidelines for Healthy Communities in 2022. When they receive reports related to their community, Mods must take action to moderate by removing content and/or escalating to Reddit Admins for review. Mods 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 also set up and configure AutoModerator to help moderate their communities. Automoderator is a tool that 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.). From January to June 2023, content removals initiated by mods accounted for 49.5% of all content removed on Reddit. Notably, mod removals can be based on any reason specific to the rules of a given community, and are not necessarily an indication of content being in violation of Reddit's Content Policy. Close to 72% of content removal actions by mods were the result of proactive Automod removals. While Automod removals clearly accounted for the majority of mod removal actions, the total volume of Automod removals decreased by 5.6% compared to the last half of 2022. This is also reflected in a roughly 3% decrease in the total volume of mod removals during that period. 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 Mods or Admins, such as a post or a comment, a chat or a private message, an award, an ad, other users, a community, a community suspected from ban evasion, an abuse of the report system, or other issues. Users can also report Mods if they believe that they violate Reddit's Content Policy or Moderator Code of Conduct. Each user may also upvote or downvote a piece of content. Sufficient numbers of downvotes result in the downranking or hiding of the content. In addition, users who get downvoted a lot will lose 'karma' (a user score that represents how much a user contributes to the Reddit community) and may be unable to access certain subreddits that require a minimum level of karma points. Reddit is not a member of the GIFCT, but participated in the GIFCT's Hash Sharing Consortium, employing automated detection methods against this hash set. However, in late 2022, the GIFCT decided to restrict hashing access to GIFCT members and cut Reddit off from the database, despite the company's strong objection. Reddit also has its own hash set based on content it finds and removes from the platform, and uses this hash set to proactively find and remove violating content. Additionally, Reddit participates in Tech Against Terrorism's TCAP alert system, which gives the company an automated notice when a terrorist-affiliated URL appears on the platform. Reddit has dedicated reporting channels to receive and promptly address reports from government officials and law enforcement agencies related to designated terrorist content, as outlined in Reddit's Guidelines for Law Enforcement
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 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 subreddit may be quarantined, face other restrictions (such as the loss of certain functionalities), or be banned, depending on the nature, scale, and seriousness of the violation.

7. Does the service issue transparency reports (TRs) on TVEC?	Yes. Reddit issues transparency reports that features 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 (1H 2023), Reddit specifically reported that out of the total amount of violent posts and comments removed (26,148 pieces of content), there were 173 pieces of terrorist content (Reddit, 2023).
8. What information/fields of data are included in the TRs?	The 1H 2023 report discloses one TVEC-specific metric and features a new, more detailed section on how Reddit identifies and removes terrorist content. Reddit's transparency report includes the following metrics: **Content on Reddit** • The total number of pieces of content created on Reddit in 1H 2023, broken down by type of content (posts, comments, private messages, chats) **Content removals** • The overall number and percentage of pieces of content removed by subreddit moderators and by Reddit administrators for violations of the Content Policy • The number and percentage of content removed by moderators, broken down by removal proactively performed by a human moderator or by Automod; • The number and percentage of content removed by Admins (not including content manipulation i.e., spam) vs. content removed by Mods • The number and percentage of content removed by Admins, broken down between spam removals, other content manipulation removals, and other content policy removals • The number and percentage of user reports (posts, comments, private messages), broken down between actionable reports and other reports (duplicate and already actioned reports, or the reported content did not violate Reddit's rules) • Number of Community Interference reports (Mods flagging to Admins instances within their own communities where a user or group of users were

	engaging in community interference), and percentage of actionable interference cases **Content types** • The overall number of posts & comments created • The number and percentage of posts & comments reported/flagged to Admins v. the number and percentage of those that resulted in removal • The percentage of reported/flagged posts and comments, broken down by source of detection (user reports v. flagged by Reddit automation) • The number of posts & comments removed, broken down by Content Policy violation (among which 'violent content'), by source of detection (user reports v. Reddit automation), and percentage of actionability (reported/flagged v. removed) • The number of potential violations of the Moderator Code of Conduct, and percentage that resulted in subreddit bans • The number of quarantined subreddits • The overall number of subreddits created v. subreddits banned (including and excluding spam-related bans) • The number of subreddits banned by categories of removal reason (including 'violent content'), and percentage of increase/decrease compared with the previous year • The overall number of private messages sent • The number and percentage of private messages reported/flagged to admins v. number and percentage of those that resulted in removal • The number of reported private messages broken down by source of detection (user reports v. flagged by Reddit automation) • The number of private messages removed, broken down by Content Policy violation (among which 'violent content'), and by source of detection (user reports v. Reddit automation) • The overall number of chat messages sent v. the number and percentage of chat messages

	reported/flagged to Admins v. the number and percentage of those that resulted in account-level sanctions • The percentage of chat messages by source of report/flag (user reports v. flagged by Reddit automation) **User accounts** • The number of accounts reported or flagged to admins v. the number and percentage of those that resulted in an action taken ('actionability') • The number of accounts sanctioned by Admins, broken down by Content Policy violations (among which 'violent content' and not including content manipulation or spam), and by type of suspension (temporary v. permanent) • The number and percentage of manual v. automated Admin actions (including content manipulation and spam) • The number of unique users who received at least one temporary/permanent ban by Mods, broken down by type of ban (temporary v. permanent ban) and by source of ban (banned by bots v. banned manually) **Appeals** • The total number of appeals received, and percentage of appeals that resulted in reversal of action • The number of appeals broken down by Content Policy violation, and percentage of appeals that resulted in a reversal of action **Legal removals** • Government and law enforcement content removal requests broken down by country (number of requests, % of requests complied with or in part, number of content or communities identified in requests, % of content or communities removed for Content Policy violations, % of content or communities restricted in requesting country, % for which no action was taken) • Content types & communities identified in government and law enforcement removal requests

	• Reddit actions in response to government and law enforcement removal requests • CSAM removals, broken down by report sources (image hash, video hash, or user report/other detection method) • Terrorist content removals, broken down by report sources (automated flagging or other source) • Private party legal removal requests, broken down by country • Reddit actions in response to private party legal removal requests • NetzDG removal requests
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Reddit's transparency reports are compiled from the company's own internal data and records, using standard data science methodologies.
10. Frequency/timing with which TRs are issued	From 2014 to 2022, Reddit issued transparency reports on an annual cadence. In the second half of 2023, the company transitioned to issuing the reports on an increased frequency of every 6 months.
11. Has this service been used to post TVEC?	Yes. The footage of the Christchurch attack was made available in at least two of Reddit's communities ('/rWatchpeopledie' and 'r/Gore'). (Hatmaker, 2019) This led to Reddit administrators permanently banning the subreddits in question from the site. Reddit's transparency reports document instances of TVEC. See Section 7 above.
12. Main changes since last Report	• Reddit now specifically defines terrorist content as "propaganda material posted by terrorists or designated terrorist organisations and their supporters, expressions of affiliation or support for terrorists or designated terrorist organisations, glorification of terrorist acts or content that solicits or incites a person or group to participate, commit, or contribute to terrorist activities." • In September 2022, Reddit introduced the Moderator Code of Conduct, which replaced the Moderator Guidelines for Healthy Communities. The new Code enables users to flag to admins instances in which moderators and/or subreddits might be in violation of the Code's rules. • In 2023, Reddit shifted from an annual to a semi-annual cadence for its transparency reports, and now releases the reports covering 6-month periods.

	• Reddit's transparency report includes more granular metrics (e.g., mod-specific metrics under the new Moderator Code of Conduct), including a new, separate section on terrorist content removals that breaks down how the terrorist content was detected (automation or other).

20. Dailymotion

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	Dailymotion defines terrorist content as 'any content that advocates or promotes violent extremist and/or terrorist organisations, individuals, or acts.' In order to identify terrorist content, Dailymotion relies on official consolidated lists of violent extremist organisations published by the UN and the European Union. Dailymotion's Prohibited Content Policy (Dailymotion, 2023), under the section titled 'Terrorist content' prohibits users from posting any content (including but not limited to): • Provoking the commission of terrorist offences or violent extremist acts or glorifying such offences; • Encouraging participation in terrorist offences; • Praising, glorifying, and/or supporting the acts of violent extremist and/or terrorist individuals or groups • Supporting terrorist and/or violent extremist ideologies; • Encouraging people to join violent extremist and/or terrorist organisations; • Providing instructions on methods or techniques for the commission of terrorist offences; • Containing images, sounds or symbols (such as names, logos, flags, slogans, uniforms, gestures, pictures, or other objects) intended to depict violent extremist and/or terrorist organisations or individuals; • Emanating from criminal, violent, extremist, or terrorist organisations; • Depicting one or more hostages; or • Posted online with the aim of soliciting, threatening, or intimidating people on behalf of a criminal, violent, extremist, or terrorist organisation. Likewise, under the section titled 'Shocking, malicious, violent or dangerous content', Dailymotion strictly prohibits any macabre, sadistic, threatening, violent, or harmful content as well as any content intended to shock or inspire disgust. Dailymotion also prohibits any content that is dangerous and/or aims to encourage, normalise, or glorify violence, suffering, death, or any

	situation that may cause injury. In particular, it is forbidden for users to post online any content:
	
	• Likely to endanger others and in particular any content encouraging, promoting, or glorifying:
	  ○ Hoaxes or dangerous challenges (strangulation challenges, scarfing, etc.);
	  ○ Suicide, self-harm and/or eating disorders;
	  ○ And in general, all practices that may harm the physical or mental integrity of the individual
	• Using excessively vulgar or offensive language;
	• Offering the sale, purchase or trade of products and services that are prohibited for sale (firearms, explosives, drugs etc.);
	• Infringing on the regulations applicable to the sale, purchase, canvassing, or trading of products and services that are regulated or prohibited for sale (firearms, explosives, ammunition, drugs, alcohol etc.);
	• Containing explicit or gruesome images of violence or serious physical harm (torture, fighting, slitting of throats, dismemberment, mutilation, human remains etc.);
	• Praise and/or incitement to join a cult or a sectarian movement;
	• Inciting violence, including any speech or action that encourages or incites the commission of violent acts or the use of force to resolve a conflict or achieve objectives such as calls to action;
	• Promoting, encouraging or providing instructions on criminal activities, violence against persons, animals or property.
	
	In addition, under the section titled 'Hateful content', Dailymotion prohibits any content which contains hateful behaviour and/or which undermines human dignity. Thus, Dailymotion strictly prohibits any content, behaviour, or statements which aim to attack, insult, threaten, or incite to hatred, violence or discrimination against a person or a group of persons on the basis of attributes or personal characteristics, real or supposed, such as: race, ethnicity, nationality, religion, caste, language, sexual orientation, sex/gender, gender identity, serious illness, disability, migratory status, and eating disorders. Such hateful content may take the form of outrageous expression, terms of contempt, invectives, stereotyping, attempts to offend, stigmatization, dehumanisation, declarations of inferiority, expressions of disgust, insults, marginalisation, or provocations of hatred, violence, or discrimination against individuals or groups of individuals targeted because of their personal attributes or

	characteristics. In particular, it is forbidden for Users to post any content: • Encouraging violence against persons or groups of persons on the basis of one or more of the attributes listed above; • Relaying hateful, insulting or humiliating speech or acts targeting, for example, the religious affiliation, gender, skin colour, sexual orientation, or disability of one or more groups of people; • Dehumanising and/or demeaning to a person or group of persons, including implying that they are inferior; • Abusing or defaming a person or group of persons on the basis of personal attribute(s) or characteristics; • Encouraging, supporting or glorifying hateful ideologies (such as antisemitism, racism, white supremacism, anti-LGBTQIA+ activism…) or proclaiming the superiority and/or dominance of one group of people over others; • Defending conspiracy theories that a person or group of people is dishonest or malicious; • Using stereotypes based on personal attributes or characteristics which incite hatred; • Encouraging, supporting, or promoting conversion therapy or any other practice aimed at changing or repressing a person's actual or perceived sexual orientation or gender identity; • Undermining human dignity; or • Contributing to the exploitation of people (such as content depicting forced labour, domestic servitude, human trafficking for sexual exploitation, slavery and/or pimping). Furthermore, under the section titled 'Content glorifying, normalising or denying a crime', Dailymotion strictly prohibits any content legitimising, glorifying, normalising, or denying violent events or international crimes committed against a population, a nation, and ethnic, racial, religious or any other group on the basis of personal attributes or characteristics. Dailymotion also prohibits any content that advocates crimes against persons or property. In particular, users are forbidden to post online any content: • Advocating or denying, minimising and/or trivialising crimes of genocide committed against the Jewish community during the Second World War; • Advocating war crimes, crimes against humanity, crimes of enslavement or exploitation of a person reduced to slavery, or crimes of collaboration with the enemy, even

	if these crimes have not led to the conviction of their perpetrators; • Denying, minimising, and/or trivialising, including in disguised or doubtful form or by way of insinuation, any other genocide against humanity, the crime of enslavement or exploitation of a person reduced to slavery or war crimes which have given rise to a criminal conviction by a French or international court • Advocating intentional attacks on life or personal integrity and/or sexual assault; or • Advocating aggravated robbery, extortion, destruction, degradation, and/or voluntary deterioration which may endanger people (vandalism, fire, etc.). Dailymotion adds that some content representing or evoking one of the situations listed above can be exceptionally maintained on the Dailymotion Service notably because of its manifestly educational, documentary, scientific or artistic context. However, Dailymotion reserves the right to reduce the visibility of such content.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://legal.dailymotion.com/en/terms-of-use/, and https://faq.dailymotion.com/hc/en-us/articles/360015770319-What-are-the-prohibited-content-on-Dailymotion-
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No. Dailymotion's rules apply to all content all features of Dailymotion including live broadcasts.
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 content is found to violate the rules detailed in its Terms and Conditions and Prohibited Content Policy, Dailymotion removes the content from Dailymotion. When content does not violate these rules but features inappropriate scenes for sensitive audiences, Dailymotion adds a restriction for sensitive content. When content is found to violate the laws of a territory and/or a country without infringing Dailymotion's internal policies, it will, where possible, restrict access to the content in the territory and/or the country where it is deemed illegal.
4.1 Notifications of removals or other enforcement decisions	Dailymotion notifies users of moderation decisions.
4.2 Appeal processes against removals or other enforcement decisions	Dailymotion has established an internal appeal mechanism for the decisions taken by its moderation team. This internal appeal system allows:

	• The author of a report to contest the moderation decision taken following his/her report; and • The User who has put content online that has been the subject of a moderation action to contest the corresponding moderation decision. This internal appeal mechanism is available by contacting Dailymotion's support teams or, in certain cases, via a URL link indicated in the e-mails sent by the moderation team to the author of the report or to the User who posted the content online. This URL link redirects to an appeal form which must be filled in by providing all the mandatory information required to process such appeal. This appeal system is open for a period of six months from the moderation decision. If the moderation team considers that the content subject to the appeal is not illicit, nor incompatible with this Prohibited Content Policy, the content concerned will be maintained or made accessible and visible again on the Dailymotion Service. On the contrary, if the moderation team considers that the content that is subject of the appeal is illegal or incompatible with this Prohibited Content Policy, a moderation action will be applied to the relevant content. The author of the appeal will be informed of the outcome of his/her appeal as soon as possible via email sent to the address given on the appeal form. The uploader of the content will also be informed and can appeal the moderation decision as well.
5. Means of identifying TVEC (for example, monitoring algorithms, user-generated, human (staff) reviewers, hash-sharing/URL sharing database)	Dailymotion uses a combination of automated tools and human review to identify TVEC on its platform: • User report mechanism: Dailymotion encourages its users and visitors to report any terrorist or violent extremist content through the user report tool available besides each video. It is not necessary to have a Dailymotion account to report content. All reported content is reviewed 24 hours a day, 7 days a week by the moderation team. • Automated tools: ○ Dynamic list of keywords and short sentences: Dailymotion has put in place a dynamic list of keywords and short sentences enabling the automatic detection of some content and, if necessary, preventing the upload of illegal content. Thus, when uploading content on Dailymotion, if the title or the description of the video contains a keyword or short sentence included in the dynamic list, the content will be automatically flagged and pushed for review to the moderation team. This list is updated

	automatically through a machine learning algorithm based on our previous moderation decisions. ○ Fingerprinting technology: In the event of the receipt of a removal order related to terrorist or violent extremist content issued by a competent authority of a Member State such as law enforcement agencies (LEA) or government, Dailymotion will send the flagged content through the INA ('Institut National de l'Audiovisuel') technology to prevent its reappearance on the platform. Each time a video uploaded on Dailymotion matches a fingerprint in the video fingerprint databases of INA, it is automatically removed. ○ Hash technology: Dailymotion has developed an in-house hashing technology to prevent the reappearance of identified terrorist or violent extremist content. Upon the identification of such content, Dailymotion removes it promptly and creates a digital fingerprint of its entire file so that the exact same content will not be available on Dailymotion again. Dailymotion joined the GIFCT, and does not participate in December 2023.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	In case of violation of its Prohibited Content Policy, Dailymotion can apply the following moderation actions: • **Removal of content** (moderation action affecting the availability of prohibited content): any prohibited content will be removed from the Dailymotion Service as soon as Dailymotion is effectively aware of it. • **Restriction of the visibility of a video** (moderation action affecting the visibility of prohibited content): Dailymotion may decide to restrict the visibility of a video. This moderation action will have the effect of placing a restricted access on the video. Such videos cannot (i) be viewed by visitors having the restricted mode activated, (ii) appear in the results of their searches on the Dailymotion platform, or (iii) be monetised nor recommended by Dailymotion. • **Application of a geo-blocking measure** (moderation action affecting the accessibility of content): If the concerned content violates a specific law/regulation in a given country/territory, Dailymotion can implement geo-blocking measures. These measures allow the content to be made unavailable in a specific country or territories

	without altering its accessibility in other countries/territories. • **Video demonetisation** (moderation action affecting the monetisation of content): Dailymotion reserves the right to demonetise video content that is not suitable for advertising purposes.
7. Does the service issue transparency reports (TRs) on TVEC?	Yes. Dailymotion published its first 'transparency report on tackling terrorist and violent extremist content' in 2023, covering January to December 2022 (Dailymotion, 2023).

8. What information/fields of data are included in the TRs?	Dailymotion's TR includes the following information:
	
	Removal orders:
	
	• Number of content removed following removal orders
	• Content which access has been disabled following removal orders
	• Number of removal orders to which Dailymotion could not comply with:
	  o On grounds not attributable to Dailymotion
	  o On grounds attributable to the removal orders
	
	Specific measures (following a decision under Regulation (EU) 2021/784 or 'TCO'):
	
	• Terrorist content removed following specific measures
	• Terrorist content which access has been disabled following specific measures
	
	TVEC removed by Dailymotion
	
	• Number of content removed following a detection by Dailymotion
	• Number and percentage of user reports received flagging potential TVEC
	• Number and percentage of content removed following a user report
	
	Complaints and review proceedings
	
	• Complaints brought by a content provider concerning a removal or disabling following specific measures:
	  o Number of complaints
	  o Outcome of the complaints
	• Administrative or judicial review proceedings brought by Dailymotion
	  o Number of review proceedings
	  o Outcome of the review proceedings
	• Number of cases in which the content was reinstated

	○ Following an administrative or judicial review proceedings ○ Following a complaint by the content provider For the period July – December 2022, Dailymotion did not receive any removal orders or decisions under the TCO. Therefore, no information is provided in these sections. The last section on 'Complaints and review proceedings' does not contain any data either.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	The TR provides the following explanations: • Removal orders: Removal orders from competent authorities of any Member State, such as law enforcement agencies (LEA) or government, requiring Dailymotion to remove or to disable access to terrorist content in Member States. • Specific measures: According to Regulation (EU) 2021/784 (or 'TCO'), a hosting service provider that is objectively exposed to terrorist content can receive a decision from the competent authority of the Member State of its main establishment. Once notified, the hosting service provider shall take specific measures to protect its services against the dissemination to the public of terrorist content.
10. Frequency/timing with which TRs are issued	TRs are expected to be published on an annual basis.
11. Has this service been used to post TVEC?	Yes. See section 8 above. For instance, in 2018, a bomb-making video linked to pro-ISIS Telegram channels and encouraging terrorist attacks was found on Dailymotion (Counter Extremism Project, 2018).
12. Main changes since last Report	Dailymotion was not included in previous Reports.

21. X

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	In 2023, Twitter was acquired by X Corp and changed its name to X. In parallel with this rebrand, X has been updating some of its policies and processes, and adjusting to the current risk landscape and regulations. In April 2023, X updated its policies and now uses different categories to classify violating content. As previously, there is no specific definition of TVEC in the 'Safety' section of the 'X Rules'. The category 'Terrorism and Violent Extremism' does

	not exist anymore and TVEC now falls under a broader category titled ‘Violent & Hateful Entities’. **Violent and Hateful Entities:** Under this policy, X states that there is no place on its platform for violent and hateful entities, including (but not limited to) terrorist organisations, violent extremist groups, perpetrators of violent attacks, or individuals who affiliate with and promote their illicit activities. The violence and hate these entities engage in and/or promote jeopardise the physical safety of those targeted. Users may not threaten terrorism and/or violent extremism, nor promote and support violent and hateful entities. Violent entities are those that deliberately target humans or essential infrastructure with physical violence and/or violent rhetoric as a means to further their cause. These include terrorist organisations, violent extremist groups, and perpetrators of violent attacks. Hateful entities are those that have systematically and intentionally promoted, supported and/or advocated for hateful conduct, which includes promoting violence or engaging in targeted harassment towards a protected category. Examples of the types of content that violate this policy include, but are not limited to, doing the following on behalf of, indirectly, or directly for a violent or hateful entity: • Engaging in or promoting violent acts • Recruiting, or providing or distributing services (such as media/propaganda) to further stated goals X may make limited exceptions for violent and hateful entities if it can determine the following: • They have reformed or denounced their violence and/or hate-based purpose. • They are currently engaged in a peaceful resolution process. • They are state or governmental entities, including those that have representatives elected to public office. In this case, X may choose to limit visibility of the content, instead of removing it. Additionally, any discussions of violent and hateful entities for clearly educational, documentary, and/or newsworthy purposes is not a violation of this policy.

	Perpetrators of violent attacks: In 2023, X released a new policy specific to 'Perpetrators of violent attacks'. It states that it will remove any accounts maintained by individual perpetrators of terrorist, violent extremist, or mass violent attacks, as well as any accounts glorifying the perpetrator(s), or dedicated to sharing manifestos and/or third-party links where related content is hosted. X does not require that a person have been confirmed as members of terrorist organisations or other violent and hateful entities, nor that they have any official affiliation with any group, organisation, or ideology, for us to enforce on content. X acknowledges that hateful and discriminatory views promoted in content produced by perpetrators are harmful for society and their dissemination should be limited in order to prevent perpetrators from publicising their message. As a result X may remove Posts that include manifestos or other similar material produced by perpetrators, even if the context is not abusive. X may also remove posts disseminating manifestos or other content produced by perpetrators, even if the context is not abusive. However, X may allow newsworthy content if it does not: • Convey suggestions about how to arm oneself and choose targets; • Share hateful slogans, symbols, memes, and/or hateful conspiracy theories; • Outline the perpetrator's ideology, tactical choices, and/or plan of attack. Furthermore, X defines a manifesto as a statement by a perpetrator outlining their motivation, views, or intent to engage in a violent attack. A manifesto can be in the form of a written document, social media post, audio recording, video, external link, or letter or other forms of content. It may be shared in the aftermath, or at any period before a violent attack. A manifesto can be linked to the event through a statement of warning or intent. **Violent Speech:** Under its 'Violent Speech' policy, users are not allowed to threaten, incite, glorify, or express desire for violence or harm. • Violent Threats: Users may not threaten to inflict physical harm on others, which includes (but is not limited to) threatening to kill, torture, sexually assault,

	or otherwise hurt someone. This also includes threatening to damage civilian homes and shelters, or infrastructure that is essential to daily, civic, or business activities. • Wishes of Harm: Users may not wish, hope, or express desire for harm. This includes (but is not limited to) hoping for others to die, suffer illnesses, tragic incidents, or experience other physically harmful consequences. • Incitement of Violence: Users may not incite, promote, or encourage others to commit acts of violence or harm, which includes (but is not limited to) encouraging others to hurt themselves or inciting others to commit atrocity crimes including crimes against humanity, war crimes or genocide. This also includes using coded language (often referred to as "dog whistles") to indirectly incite violence. • Glorification of Violence: Users may not glorify, praise, or celebrate acts of violence where harm occurred, which includes (but is not limited to) expressing gratitude that someone experienced physical harm or praising Violent entities and Perpetrators of Violent Attacks. X allows expressions of violent speech when there is no clear abusive or violent context, such as (but not limited to) hyperbolic and consensual speech between friends, or during discussion of video games and sporting events. X also allows certain cases of figures of speech, satire, or artistic expression when the context is expressing a viewpoint rather than instigating actionable violence or harm. **Hateful Conduct:** Under its 'Hateful Conduct', policy, X does not allow users to directly attack other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. Action will be taken against reports of users targeting individuals or groups of people with the following behaviours: • <u>Hateful references:</u> X 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: X prohibits inciting behaviour that targets individuals or groups of people belonging to protected categories. This includes:
	  ○ inciting fear or spreading 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.”
	  ○ inciting 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!“
	  ○ inciting 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.
	• Slurs and Tropes: X prohibits targeting others with repeated slurs, tropes or other content that intends to degrade or reinforce negative or harmful stereotypes about a protected category. In some cases, such as (but not limited to) severe, repetitive usage of slurs, or racist/sexist tropes where the context is to harass or intimidate others, we may require post removal. In other cases, such as (but not limited to) moderate, isolated usage where the context is to harass or intimidate others, we may limit post visibility.
	• Dehumanisation: X prohibits the dehumanisation of a group of people based on their religion, caste, age, disability, serious disease, national origin, race, ethnicity, gender, gender identity, or sexual orientation.
	• Hateful Imagery: X considers hateful imagery 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. ○ Media depicting hateful imagery is not permitted within live video, account bio, profile or header images. All other instances must be marked as sensitive media. Additionally, sending an individual unsolicited hateful imagery is a violation of this policy. • Hateful Profiles: Users may not use hateful images or symbols in their profile image or profile header. Also, users may not use their username, display name, or profile bio to engage in abusive behaviour, such as targeted harassment or expressing hate towards a person, group, or protected category. **Sensitive media:** Under this policy, users may not post media that is graphic or share violent or adult nudity and sexual behaviour within live video or in profile header, List banner images, or Community cover photos. Examples of prohibited 'graphic content' include: ○ violent crimes or accidents; ○ physical fights; ○ physical child abuse; ○ bodily fluids including blood, faeces, semen etc.; ○ serious physical harm, including visible wounds; and ○ severely injured or mutilated animals. **Illegal or Certain Regulated Goods or Services:** Under this policy, X does not allow selling, buying, or facilitating transactions in illegal goods or services, as well as certain types of regulated goods or services. This includes weapons, such as firearms, ammunition, and explosives, and

	instructions on making weapons (e.g. bombs, 3D printed guns, etc.)
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://help.twitter.com/en/rules-and-policies/x-rules
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No. The X Rules apply to live 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?	When determining whether to take enforcement action, X 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 X policies; • the severity of the violation; • the content may be a topic of legitimate public interest (X, 2023). X has a range of enforcement options that it may exercise when a user violates the X Rules (X, 2023): • **Tweet-level enforcement:** X takes action at the Tweet level when a specific Tweet violates the X Rules, including Tweets that share or reproduce other Tweets by posting screenshots, quote-Tweeting, or sharing Tweet URLs that violate its Rules. ○ Limiting Tweet visibility: Where appropriate, X will restrict the reach of Tweets that violate its policies and create a negative experience for other users by making the Tweet less discoverable on X. This can include: ▪ Excluding the Tweet from search results, trends, and recommended notifications ▪ Removing the Tweet from the For you and Following timelines ▪ Restricting the Tweet's discoverability to the author's profile ▪ Downranking the Tweet in replies

	▪ Restricting Likes, replies, Retweets, Quote Tweets, bookmarks, share, pin to profile, or Edit Tweet o Excluding the Tweet from having ads adjacent to it: Starting in April 2023, Tweets identified as violating the Rules will begin to receive labels informing both Tweet authors and viewers that X limited the Tweet's visibility. o Requiring Tweet removal: When X determines that a Tweet violated the X Rules and the violation is severe enough to warrant Tweet removal, X will require the violator to remove it before they can Tweet again. o Labelling a Tweet: If X determines a Tweet contains misleading or disputed information per its policies that could potentially lead to harm, it may add a label to the content to provide context and additional information to users. In these cases, Community Notes may also be visible on Tweets to provide additional context. o Notice of public interest exception: X may determine that it is in the public interest for a Tweet that would otherwise be in violation of the rules to remain accessible on its service. When this occurs, X will place the Tweet behind a notice and limit its visibility. • **Direct message-level enforcement:** In a private direct message conversation, when a participant reports the other person, X will stop the violator from sending messages to the person who reported them. The conversation will also be removed from the reporter's inbox. However, if the reporter decides to continue to send Direct Messages to the violator, the conversation will resume. • **Account-level enforcement:** X takes action to suspend an account if it determines that a user has engaged in repeated violations of its policies and/or violated specific policies that cause significant risk to X (i.e. posting illegal content, attempts to manipulate our platform or spam users, using our platform to incite violence, etc.) or pose a threat to our users (fraud, user privacy violations, violent threats, targeted harassment, etc.). o Placing an account in read-only mode: X may temporarily limit an account's ability to Tweet, Retweet, or Like. The user can read

	their timelines and will only be able to send Direct Messages to their followers. ○ Verifying account ownership: X may require an account owner to verify ownership with a phone number or email address. 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. • **Actions against non-violating content** ○ Placing a Tweet behind a notice: X may place some forms of sensitive media like adult content or graphic violence behind an interstitial advising viewers to be aware that they will see sensitive media if they click through. ○ Withholding a Tweet based on age: X restricts views of specific forms of sensitive media such as adult content for viewers who are under 18 or viewers who do not include a birth date on their profile. ○ Withholding a Tweet or account in a country: X may withhold access to certain content in a particular country if it receives a valid and properly scoped request from an authorised entity in that country.
4.1 Notifications of removals or other enforcement decisions	Notifications take place typically when X requests a user to modify their behaviour and be in compliance with its rules (requiring media or profile edits), or in case of permanent account suspension. When X 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 (X, 2023).
4.2 Appeal processes against removals or other enforcement decisions	Users can appeal a locked or suspended account if they believe X made an error. Upon appeal, if it is found that a suspension is valid, X responds to the appeal with information on the policy that the account has violated. Users can also submit an appeal for a posy removal, or after one of their posts have been excluded from having ads adjacent to it.
5. Means of identifying TVEC (for example, monitoring algorithms, user-generated, human (staff) reviewers, hash-sharing/URL sharing database)	X has three primary ways of detecting content that may violate its rules. 1. User reporting: Anyone can report potential violations of X's policies, whether they have an X account or not. People have the ability to

	report a post, List, or profile; specific content in a Moment; a Space or person in a Space; or a product. Moderators review the reports and decide whether the content in fact violates X's rules. A DSA reporting form has also been made available and is meant to allow the reporting of illegal content within the EU on the basis of the applicable Union law or national law in compliance with Union law. 2. Proactive content-based detections X also uses internal, proprietary tools to detect violations of the 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 X utilises internal, proprietary tools to detect violations of the Rules, including the posting of TVEC, based on the behaviour exhibited that can be associated with terrorist organisations. X is member of the GIFCT.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	In 2022, X announced the creation of a content moderation council with widely diverse view points and that no major content decisions or account reinstatements would happen before that council convenes (Frenkel, 2022). However, no new information has been released since then. Depending on the policy violated, X will apply the following sanctions: • Violent and Hateful Entities: Immediate and permanent suspension of the violating account. • Violent Speech: In most cases, X will immediately and permanently suspend any account that violates this policy. For less severe violations, it may instead temporarily lock users out of their account before they can post again. In rare cases, it may make the violative content less visible by restricting its reach on X. However, if you continue to violate this policy after receiving a warning, your account will be permanently suspended. We also recognise that conversations regarding certain individuals credibly accused of severe violence may prompt outrage and associated violent speech. In these limited cases, we may take less punitive measures.

	• Hateful Conduct: When determining the penalty for violating this policy, X considers a number of factors including, but not limited to the severity of the violation and an individual's previous record of rule violations. The following is a list of potential enforcement options for content that violates this policy: o Making content less visible on X by: ▪ Removing the Post from search results, in-product recommendations, trends, notifications, and home timelines ▪ Restricting the Post discoverability to the author's profile ▪ Downranking the Post in replies ▪ Restricting Likes, replies, Reposts, Quote, bookmarks, share, pin to profile, or engagement counts ▪ Excluding the Post from having ads adjacent to it o Excluding Posts and/or accounts in email or in-product recommendations. o Requiring Post removal. ▪ For example, we may ask someone to remove the violating content and serve a period of time in read-only mode before they can Post again. o Suspending accounts that violate our Hateful Profile policy. • Illegal or Certain Regulated Goods or Services: X may suspend the account, including upon first review.
7. Does the service issue transparency reports (TRs) on TVEC?	No. X used to publish semi-annual transparency reports (Twitter, 2012-2022) focused on X Rules enforcement. However, the latest report was published in 2022, covering July to December 2021, and no new TR has been published since.
8. What information/fields of data are included in the TRs?	X has not published transparency reports since 2022.
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. Previous transparency reports published by X contain information on TVEC. According to researchers focused on online extremism, activity from terrorist organisations operating on X has increased by at least 69% since its acquisition by X Corp (Naffakh, 2022).
12. Main changes since last Report	• In 2023, Twitter was acquired by X Corp and changed its name to X. With this rebrand, X has been updating some of its policies and processes. X announced that its owner, Elon Musk, and Chief Executive, Linda Yaccarino, would both oversee the trust and safety team. It is not clear yet what changes this internal reorganisation will entail for content moderation on X. • X has not published any new TR since 2022 (covering July – December 2021). • X updated its 'Safety' policies and uses new categories of violating content. o The category 'Terrorism and Violent Extremism' does not exist anymore and TVEC now falls under a broader category titled 'Violent & Hateful Entities'. o The new policy does not mention 'national and international terrorism designations' as it was the case in the previous policy. o The new policy does not mention offline activities of violent organisations anymore, whereas previously, X explained that it examined both on and off activities to determine whether a group was engaging in and/or promoting violence. o In change, the policy now allows new (limited) exceptions for violent and hateful entities, for example if they are engaged in a peace process or are elected in public office. o X also adopted a new policy specific to 'Perpetrators of violent attacks', under which it removes any account maintained by individual perpetrators of terrorist and violent extremist attacks. • X modified its enforcement policy as follows: o Added: excluding the Tweet from having ads adjacent to it; actions against non-violating content (placing a Tweet behind a notice; withholding a Tweet based on age; and withholding a Tweet or account in a country).

	○ Removed: requiring media or profile edits (at the account-level) • In 2022, X announced the creation of a content moderation council but not further information has been released yet.

22. Bilibili

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, Bilibili's Community Rules prohibit the publishing of 'violent, harmful, or dangerous content', which is defined as content that intentionally incites violence or depicts or encourages others to participate in illegal activities or activities that may cause injury or death. Under this policy, Bilibili prohibits: • Content that engages with any of the following activities: terrorist activities, organised criminal activity; videos that express support for the above-mentioned violent or criminal activities, graphic content; content that contains names, symbols, slogans, or other symbolic material intended to represent dangerous individuals or organisations and their activities; content that praises or supports the leaders of the aforementioned organisations or condones their violent activities. • Videos depicting violence, gore, or other gruesome content. If the video is from a news report or documentary, users must provide a description that clearly explains the background information for the content to avoid causing unnecessary distress to viewers and other users. • Videos that incite others to commit violent acts. This includes content that instigates acts of violence or threatens others with extreme violence. • Content that encourages dangerous or illegal activities including the purchase, sale, and production of bombs, drugs, and other prohibited items, or other actions that may result in serious harm.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.bilibili.tv/en/user-agreement and https://www.bilibili.tv/marketing/protocal/communityrules_en.html
3. Are there specific provisions applicable to livestreamed content	No.

in 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?	Bilibili's Community Rules state that Bilibili reserves the right (but does not have the obligation) to review and oversee the content published by users and take any necessary actions. These actions include, but are not limited to, deleting or blocking the relevant content without notifying users, revoking accounts, restricting, suspending, or terminating access to all or parts of Bilibili's services, or holding users legally accountable for violating the Community Rules.
4.1 Notifications of removals or other enforcement decisions	No. Notifications of enforcement actions are at Bilibili's discretion.
4.2 Appeal processes against removals or other enforcement decisions	No information is provided.
5. Means of identifying TVEC (for example, monitoring algorithms, user-generated, human (staff) reviewers, hash-sharing/URL sharing database)	Users can report content if they believe that it violates Bilibili's policies. In particular, users can report video submissions, chat messages, and comments. The reporting system does not automatically remove reported content. After it receives a report, Bilibili reviews the content in accordance with its Community Rules. Should the violation be confirmed, the content will be immediately blocked, deleted, or dealt with using a different method. Reports conducted via the reporting system are done so anonymously; other users will not be able to find out the identity or account information of the reporter. Bilibili uses a combination of automated tools and human review (Bilibili, 2021): • AI-powered screening system: Bilibili uses automatic comparison, labelling and screening of pirated, illegal, inappropriate content and other content that violates Community Rules. Bilibili launched its self-developed AI system in 2021, the 'Avalon Community Self-Purification System'. It explains that it is designed to analyse users' intentions and behaviours and intercept negative content while recommending quality bullet chats and comments. In 2021, the Avalon System automatically processed over 720,000 pieces of negative content per day. • Content audit team: All content, particularly flagged content by the AI screening system, is manually checked. As of the end of 2022, Bilibili's content audit team totalled 3,874 employees. Bilibili provides training for content auditors and conduct assessments through its online platform.

	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 L. , 2019). Furthermore, Bilibili relies on its community to monitor content on its platform. The 'Discipline Committee' is composed of voluntary users in charge of screening content, maintaining community order, and providing weekly feedback on community issues. Users must submit an application to become a member and they must respect certain conditions (no violations within the last 90 days and real-name authentication) (Bilibili, 2023). However, no specific information is provided on the moderation tools available to the Committee's members. Bilibili is not a member of the GIFCT.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	In response to any actions that violate the Community Rules, Bilibili and Bilibili's appointed community managers reserve the right to take measures such as blocking content, removing content, banning accounts, issuing warnings, and other appropriate measures. Violations may also, if required by applicable laws and regulations, be reported to the relevant regulatory bodies (Bilibili, 2022).
7. Does the service issue transparency reports (TRs) on TVEC?	No. Bilibili issues annual 'Environmental, Social and Governance Reports' that feature a section on 'Content Safety Assurance'. It contains general information on Bilibili's content audit mechanisms but there is no specific data on TVEC (Bilibili, 2022). In addition, Bilibili publishes information on user bans on a dashboard titled the 'black room'. Each case is documented with explanations on the case, the reason justifying the ban and its duration (Bilibili, 2023). However, there is no specific data 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.
12. Main changes since last Report	Bilibili was not included in previous Reports.

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 Policies explicitly ban TVEC and associated activities. In 2022, LinkedIn updated its Professional Community Policies that now include new categories of violating content with more precise information. TVEC is now included in the category titled 'Dangerous organisations and individuals' (whereas before, TVEC was not included in a specific policy and LinkedIn required members to not post terrorist content nor promote terrorism and violent extremism). In substance, the 'Dangerous organisations and individuals' category prohibits the same content and behaviours, but also extends to other types of dangerous organisations. LinkedIn also provides more detailed explanations and examples of what is included in each policy. **Dangerous organisations and individuals:** 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 have a LinkedIn profile. 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. LinkedIn also restrict profiles and pages associated with dangerous organisations and individuals regardless of whether they have posted violative content. Examples under this policy include: • Content that visually depicts acts of terrorism or that promotes or propagandises terrorist groups, individuals, or activities • Glorification or incitement of dangerous organisations and individuals, acts of terror or violent extremism • Recruitment for terrorist groups or other organisations that espouse violence • Profiles or pages created and maintained by or in support of terrorist organisations or terrorist individuals, including:

	○ Any non-state group that (1) identifies through its stated purpose, publications, or actions as an extremist group, (2) engages or has engaged in violence and / or the promotion of violence to further its cause, and (3) targets civilians (non-military) ○ Any individual that appears in the US Federal Bureau of Investigation's Domestic Terrorism or Most Wanted Terrorists Lists or is a Specially Designated Global Terrorist (SDGT) by the United States Department of State or the US Department of the Treasury LinkedIn adds that depictions of terrorist or violent extremist acts can sometimes raise awareness or condemn and there are instances when the context indicates that it was shared for one or both of those purposes. In these instances, members won't be penalised, but LinkedIn may take steps to label or obscure such content to protect other members who may not wish to see it. **Violent and graphic content:** 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. Examples under this policy include: • Depictions of mutilated and/or deceased persons or body parts (including scenes of crimes or accidents) • Blood, gore, and human or animal fluids or waste • Depictions of animal cruelty • Realistic depictions of severe physical violence against a human being, including: ○ Murder or attempted murder ○ Rape ○ Torture ○ Beatings ○ Mass shootings ○ Extreme brutality In cases where violent or graphic content is shared in connection with a newsworthy event, LinkedIn may label such content instead of removing it. In cases where graphic content or content that may otherwise be sensitive for some members is shared to raise awareness or condemn, LinkedIn won't penalise members for posting content in these circumstances,

	but may take steps to limit its distribution to protect other members who may not wish to see it. **Hateful and derogatory content:** 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, age, 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. Examples under this policy include: • Slurs and pejoratives used to demean others on the basis of their inherent traits • Promoting or expressing support for hate groups or ideology via symbol or otherwise • Expressions of disgust towards or superiority over people or groups on the basis of inherent traits • Misgendering or deadnaming of transgender individuals • Calls for exclusion or banishment of a specific group based on inherent traits • Wishing that members of an inherent group die or suffer as a result of violence or serious disease • Negative stereotypes, rooted in inherent traits, that reinforce harmful attitudes or behaviours against individuals or groups • Subhuman characterisations, including comparisons to insects, animals, filth, disease, and denial of existence • Holocaust denial or misappropriation of Holocaust symbology • Denial of slavery in the United States LinkedIn may label rather than remove content that evokes hateful rhetoric (including slurs) in the context of counter speech, reclamation, or members' personal experiences with racism, sexism, ableism, and other forms of prejudice or discrimination.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.linkedin.com/legal/professional-community-policies
3. Are there specific provisions applicable to livestreamed content in	Yes. In addition to having to comply with the ToS and the LinkedIn Professional Community Policies, live-streaming is a

the ToS or Community Guidelines/Standards?	limited feature on LinkedIn. Any member who wants to create and host an event must meet the following eligibility criteria: • Audience base: minimum 150 followers and/or connections • A history of abiding by LinkedIn's Professional Community Policies • Geography: LinkedIn Live is not available for members and Pages based in mainland China By creating an event or turning creator mode on, members automatically trigger a review of their profile or page by LinkedIn who then decides to allow the Live functionality or not, on the basis of this criteria (LinkedIn, 2023). There is no application form anymore. 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 members to report content that violates its Professional Community Policies. When a member 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 reviews reported content or conversation to assess whether it violates its Professional Community Policies. Violations of LinkedIn's ToS or Professional Community Polices can result in action against the account or content.
4.1 Notifications of removals or other enforcement decisions	Content removals are generally notified. LinkedIn also explains how the content violates its policies and the type of action that is being taken as a result.
4.2 Appeal processes against removals or other enforcement decisions	If an account has been restricted or content has been removed and the member believes the action was in error, the member can appeal the decision.
5. Means of identifying TVEC (for example, monitoring algorithms, user-generated, human (staff) reviewers, hash-sharing/URL sharing database)	Members are able to report content that violates LinkedIn's policies, as well as profile photos, profiles, groups, or messages. 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, LinkedIn removes such content. In addition, LinkedIn employs machine classifiers and processers to detect potential TVEC on its platform. In 2023,

	LinkedIn shared its 'Responsible AI Principles' which include providing transparency on how AI impacts people (Lawit & Xu, 2023). LinkedIn is a member of the GIFCT.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Violating LinkedIn's User Agreement and Professional Community Policies can result in action against an account or content. Depending on the severity of the violation, LinkedIn may limit the visibility of certain content, label it, or remove it entirely. Repeated violations may result in account restriction. Continued violations will result in permanent restriction from the LinkedIn platform. For certain egregious violations of LinkedIn's Professional Community Policies (e.g., child sexual abuse material, terrorism, extremely violent content, egregious sexual harassment), LinkedIn may permanently restrict an account after a single violation. Content that would normally violate the letter of LinkedIn's policies may be allowed in cases where the content is being shared for awareness or to condemn. In these cases, it may label and obscure the content for members who may find this content sensitive or disturbing, or otherwise do not want to view it. However, LinkedIn won't remove the content or penalise the author for posting it (LinkedIn, 2023).
7. Does the service issue transparency reports (TRs) on TVEC?	Not specifically. LinkedIn issues semi-annual transparency reports 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 gory, gruesome, or disturbingly shocking" (LinkedIn, 2023) and thus is broader than TVEC alone. The latest report is available at https://about.linkedin.com/transparency/community-report LinkedIn also publishes an annual report under Regulation (EU) 2021/784 on addressing the dissemination of terrorist content online. The latest report is available at https://www.linkedin.com/help/linkedin/answer/a1505695
8. What information/fields of data are included in the TRs?	LinkedIn's last transparency report, which covers the period July to December 2022, 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.
10. Frequency/timing with which TRs are issued	On a semi-annual basis.
11. Has this service been used to post TVEC?	Possibly. Research has shown that US-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).
12. Main changes since last Report	• In 2022, LinkedIn updated its Professional Community Policies that now include new categories of violating content, with more precise explanations and examples. TVEC is now included in the category titled 'Dangerous organisations and individuals'. This policy clearly prohibits TVEC and provides examples. Under this policy, LinkedIn explains that it also relies on the US Federal Bureau of Investigation's Domestic Terrorism or Most Wanted Terrorists Lists, and the Specially Designated Terrorist (SDGT) list by the US Department of State or the US Department of the Treasury. • LinkedIn now explains that some content may not be removed if it is posted to raise awareness or condemn terrorist and violent extremist acts; but LinkedIn may take steps to label such content to protect other users. • LinkedIn provides more precise information on its enforcement policies. For instance, in the case of TVEC, LinkedIn may permanently restrict an account after a single violation. • There is no application form to request LinkedIn Live anymore. By creating an event or turning creator mode on, members automatically trigger a review of their profile or page by LinkedIn who then decide to allow the Live functionality or not, based on a set of criteria.

24. 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 (Baidu Tieba, 2023) prohibit content and behaviour that promotes extremism, terrorism, ethic separatism, and religious fanaticism, as well as ethnic hatred, violence, and murder. Baidu Tieba also states that illegal content includes but is not limited to text, pictures, dynamic content, and videos.
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?	See section 6 below.
4.1 Notifications of removals or other enforcement decisions	The users of the content that is removed will be notified through messages sent via the platforms.
4.2 Appeal processes against removals or other enforcement decisions	Yes. When their content is removed, users are given the right to appeal. After receiving such an appeal, the corresponding content will be reviewed on a case-by-case basis, with a completion rate of 100%.
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 allows 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. Baidu, Baidu Tieba's parent company, uses a combination of automated tools and human review to moderate content. • Automatic content moderation: AI technology is leveraged to preliminary filter information and identify highly-matched illegal content samples and then remove such samples directly.

	○ Texts: A million-word illegal entries vocabulary is employed at the bottom layer to conduct word and multi-mode search and matching. The application layer model is built on various dimensions based on the natural semantic recognition technology and is then applied to monitor content involving pornography, gambling, and fraud, to ensure the recall of prohibited information. ○ Images: A million-picture prohibited image database is employed at the bottom layer to recall potential similar images by similar image recognition, face recognition, incomplete face recognition, important logo detection. The capability of recognition of local features of potentially harmful content is also enhanced to identify content involving illegal activities, pornography, violence, and terrorism, so as to ensure the recall of prohibited content. ○ Videos: Such basic capabilities as key frame extraction, ASR audio-to-text conversion, and voiceprint recognition are employed at the bottom layer, combined with video understanding capabilities such as indecent actions and scene recognition in videos, are applied to detect content involving pornography, violence and terrorism, so as to ensure the recall of prohibited content. • Manual content moderation: Manual moderation is performed on suspected harmful content that is difficult for algorithms to judge, so as to improve moderation accuracy. 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 L. , 2019). Baidu Tieba is not a member of the GIFCT.
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. Since 2020, Baidu Tieba's parent company, Baidu Inc., issues Environmental, Social and Governance Reports (Baidu, 2022). In these reports, the section on 'Content governance' features transparency reporting on Baidu's content moderation across all its services.

	The 2022 report features information on content removal, proactive detection, and appeals, but no TVEC-specific information. However, terrorism and incitement to violence are listed as examples of focuses for content governance at Baidu and may be included in the broader category titled 'Violence'.
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. Main changes since last Report	• Since 2020, Baidu Tieba's parent company, Baidu Inc., issues Environmental, Social and Governance Reports. In these reports, the section on 'Content governance' features transparency reporting on Baidu's content moderation across all its services. Although they do not provide TVEC-specific information, terrorism and incitement to violence are listed as examples of focuses for content governance at Baidu, and may be included in the broader category titled 'Violence'. • Baidu provides new information on its notification and appeal mechanisms. The users of the content that is removed are notified through messages sent via the platforms, and users are given the right to appeal. After receiving such an appeal, the corresponding content will be reviewed on a case-by-case basis, with a completion rate of 100%. • Baidu provides more detailed information on its use of AI technologies to detect illegal and harmful content, including violence and terrorism, in texts, images, and videos.

25. 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 harassing, abusive, threatening, harmful, vulgar, obscene, or offensive, or that contains pornography, nudity, or graphic or gratuitous violence, or that promotes violence, racism, discrimination, bigotry, hatred, or physical

	harm of any kind against any group or individual, or 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	Notifications are at Douban's discretion. Douban may remove or modify user content at any time for any reason, in its sole discretion, with or without notice to the relevant user.
4.2 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 L. , 2019). Douban is not a member of the GIFCT.
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.
12. Main changes since last Report	No main changes since last Report.

26. Moj

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	Moj's Content Guidelines ("Community Standards/Guidelines") (Moj, 2022) do not provide specific definition. However, they explicitly prohibit terrorism and violence including organised violence. Moj's policies are similar to ShareChat's Content Policy. **Violence:** Moj states that it prohibits violence, which includes all content that causes discomfort to users due to the goriness in the content, such as but not limited to graphical images or videos that glorify violence and suffering, or intends to incite violence, depiction of physical violence or animal cruelty. Content which promotes dangerous and illegal activities, or praises individuals, groups or leaders involved in terrorism, organised violence or criminal activities is strictly prohibited. Educative or informative content pertaining to violence may be allowed on the platform. Violent content on the platform in the form of fictional set up or martial arts may be permitted. **Illegal Activities:** Moj has zero-tolerance for content that advocates or promotes illegal activities. It prohibits content related to organised crime, criminal activities, promotion/sale/use of weapons, firearms and explosives, violence or terrorist activities.

	Users are not allowed to post content that displays tutorials or instructions or educates the users about illegal and prohibited activities including, but not limited to participating in criminal activities, making bombs or encouraging or doing or trading in drugs. Users must not use the platform to solicit or facilitate any transaction or gift involving such goods and services which are declared illegal by the Government of India. **Hate Speech and Propaganda:** Content that promotes violent behaviour against an individual or a group of individuals, intends to intimidate, target or demean any particular religion, race, caste, ethnicity, community, nationality, disability (physical or mental), diseases or gender, is prohibited. Any kind of content which produces hatred or has the intention of creating or spreading hatred or hate propaganda along the lines of including, but not limited to religion, caste, ethnicity, community, sexual orientation, or gender identity is also not allowed. Moj does not entertain content that spreads discrimination, intends to justify violence based on the above-mentioned attributes and refers to an individual or a group of individuals as inferior in any sense or with negative connotations. Moj urges users to refrain from incendiary commentary and publishing theories or hateful ideologies that may cause outrage to other users and influence them negatively. Moj may permit such content which intends to raise awareness about these issues or challenge it, subject to clear intention of posting such content on the platform. Lastly, Moj's Terms of Use (Moj, 2022) state that users shall not use the platform to share any content which is obscene, pornographic, harmful for minors, discriminatory, spreading what may be considered as hate speech, inciting any form of violence or hatred against any persons, or violates any laws of India, or is barred from being shared by any laws of India. Moj reserves the right to remove such content.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://help.mojapp.in/policies/content-policy/, and https://help.mojapp.in/policies/terms
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	Yes. Moj's Terms of Use (Moj, 2022) contain specific provisions applicable to Moj Live. The content uploaded using the Livestream feature must conform to Moj's policies. Moj reserves the right to

	immediately remove or suspend any livestream and/or take other such actions that may be relevant. If any such removal/termination/suspension action has been taken against a user, he or she may appeal the action through the in-app appeals mechanism or write to Moj at contact@sharechat.co. Moj may disable comments for livestream hosted by underage users to ensure safety and security on the platform. Furthermore, Moj encourages users to report any livestream or comment on such livestream that may be in violation of applicable laws, its Terms of Service including the Community Standards. Users may make use of the in-app reporting mechanism or write to Moj at contact@sharechat.co. Moj temporarily records and stores livestreams for a duration of 21 days or above for purposes of compliance with applicable rules and regulations to assess user-reported incidents on livestreams, and to facilitate co-operation with lawful authorities and law enforcement mechanisms. It may delete these recordings post the completion of 21 days duration, provided there are no reports on the livestream. However, it may still store the same for a longer period of time in order to cooperate with legal authorities and law enforcement mechanisms.
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?	Moj actively removes content which is not allowed and violates both its Guidelines as well as applicable Indian laws. If such content comes to its attention, it may take it down or ban user accounts. Moj also encourage users to report content that may be in violation of its Community Standards. Moj adds that the intent of the creator is important. While it understands the importance of creative freedom, it does not welcome content that intends to bring discomfort, spread what may be considered hate speech and abuse or promote violence and illegal activities.
4.1 Notifications of removals or other enforcement decisions	Moj notifies users when their content is removed and specifies the reason for removal. This helps the user understand the reason for any removal action taken by Moj.
4.2 Appeal processes against removals or other enforcement decisions	If users believe that their content has been unfairly removed, they may raise an in-app appeal via the ‘Violations’ page accessible through the Profile Settings option on the users' profile. They may also contact grievance@sharechat.co to dispute the action. Moj may reassess the content and

	determine the validity of appeals raised by the user.If users believe that their content has been unfairly removed, they may contact grievance@sharechat.co to challenge the removal. Moj may review the content again and determine if it may be reposted on the platform.
5. Means of identifying TVEC (for example, monitoring algorithms, user-generated, human (staff) reviewers, hash-sharing/URL sharing database)	Moj uses a combination of human moderators and deep learning-based Computer Vision models (commonly known as AI) to implement its Community Standards. These AI model(s) facilitate the detection of harmful content on Moj. Once flagged by the model, such content goes through human moderators, who decide the action to be taken as per various standards set forth by the Guidelines and standards assigned to the model. Moj's approach to content moderation comprises the following: • Proactive approach: Each post generated on Moj undergoes scrutiny by a series of AI model(s) to calculate a risk score. These models typically flag content based on assigned labels such as sexually explicit material, violence, or the presence of blood/weapons. Following the risk scores assigned to these predictions, content with high-risk scores is prioritised for manual review. Subsequent actions taken on flagged content and the priority for manual review are determined by the severity of the violation. For instance, 'Not Safe For Work' (NSFW) content may be prioritised higher in manual moderation and may invoke a stricter ban on the user. This manual review of flagged content not only aids in reducing false rejections but also contributes to fine tuning the model, as automated actions may be taken regarding similar posts that are in violation of the Community Standards. • User feedback: Apart from the above-mentioned flagging system, content that is in violation of Moj's Community Standards can also be reported by logged in and non-logged in users. Users can report both user profiles and individual content using the in-app reporting mechanism across multiple categories including violence, illegal activities, terrorism, hate speech, and more. In addition to the in-app reporting mechanism, users can also raise reports through other reporting mechanisms as mentioned below: o In-app Help and Support Chat: Registered/logged-in users on Moj can go to their profiles and access the Help and Support Chat and report any violation of the Community Standards;

	○ Email Support: An individual can also contact the Moj support team via email. The relevant email id is contact@sharechat.co; ○ Grievance Officer: Moj has a Grievance Officer to address platform usage concerns. The relevant email ID is grievance@sharechat.co. ○ Reactive approach: Another strategy developed at Moj is to routinely review highly viral content as such content may impact a large number of users. Every post that crosses a preset threshold of virality is reviewed manually by the human moderators. The reports are reviewed, and appropriate actions are taken by Moj's user grievance and content moderation teams, in accordance with the Community Standards and applicable rules and regulations. Moj also uses relevant content reports for bettering their AI models. Automated content moderation tools used by Moj include for example label propagation (i.e., propagating the labels from a few set of labelled examples to unlabelled examples based on their similarity with the labelled examples), active learning, semi-supervised learning, classifiers, vision models to process images and videos etc (Mandav, Parihar, Saket, Gupta, & Mukherjee, 2021). Moj is not a member of the GIFCT.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Moj takes strict and prompt action against those who violate its Community Standards. If a profile violates Moj's Community Standards, it may be suspended temporarily for a short duration (as specified below). In case of repeated breach of the Community Standards or attempt to circumvent a suspension, Moj may be compelled to take a stricter action and restrict access to the account for a period of 365 days. The banned user is prohibited from re-registering on the platform for the duration of such ban using the same credentials. Moj cooperates with lawful requests and orders from legal authorities as well as law enforcement mechanisms in accordance with applicable laws. Currently, Moj implements the following enforcement measures for content and profiles that violate its Community Standards: (i) Content takedowns; (ii) Account-level bans on platform usage and; (iii) Feature-specific bans.

	The ban durations can range from hourly bans (3- and 6-hour bans) to one day, three days, seven days, 30 days, 360/365 days, as applicable: • User-generated content ban: User is unable to post any content on the platform for the specified ban duration. • Edit profile ban: user is unable to edit any profile attributes for the specified time period. • Comment ban: user is banned from commenting on any post on the platform for the specified ban duration. • Livestream ban: User is unable to use the livestream feature (Moj, 2023).
7. Does the service issue transparency reports (TRs) on TVEC?	Yes. Moj publishes monthly transparency reports. The latest available report covers the month of September 2023 (Moj, 2023).
8. What information/fields of data are included in the TRs?	Moj's TR includes one TVEC-specific information: • Number and percentage of content and accounts reported, based on various categories as tagged by users such as 'Terrorism', 'Hate speech', 'Illegal activities', and 'Violence'. The TR includes other non-TVEC-specific information: **Law enforcement requests:** • Law enforcement requests received • Requests where user data was provided • Requests wherein a takedown/ban action was taken on the basis of violation of the community guidelines **User reports:** • Total number of user reports received • Number and percentage of user reports (content and accounts) broken down by report reasons **Enforcement actions (takedown)** • Number of content pieces taken down by categories of content **Enforcement actions (profile bans)** • Number of user-generated content bans, edit profile bans, comment bans, livestream bans, broken down

	by ban duration (hourly, 1 day, 3 days, 7 days, 30 days, 360 days)
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Yes. See section 8 above.
10. Frequency/timing with which TRs are issued	On a monthly basis.
11. Has this service been used to post TVEC?	No information is provided.
12. Main changes since last Report	Moj was not included in previous Reports.

27. 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.
	
	Quora updated its Acceptable Use Policy in 2022 and now explicitly prohibits acts of violence by terrorist organisations, whereas before it broadly prohibited 'glorifying or advocating violence'.
	
	Quora differentiates between two sets of policies: Platform Policies, that apply to the entire Quora platform, and Spaces Policies, that apply at the Space-level. Spaces are places where users can share questions, answers, and other content to collectively build knowledge and community. Both policies prohibit the same types of content.
	
	Under these policies, Quora prohibits:
	
	Harmful activities and self-harm
	
	• Threatening violence or calling for serious physical harm to an individual
	• Encouraging, glorifying, or promoting:
	  ○ Acts of physical violence towards others by civilians (excluding discussions about reasonable self-defence as well as violence involving military actions)
	  ○ Acts of physical violence towards others by other non-state actors, such as terrorist organisations
	  ○ Suicide or self-harm (including eating disorders). This includes soliciting or sharing information, strategies, or methods on how to commit suicide or self-harm

	○ Animal cruelty • Stating a specific intent to commit physical violence • Stating intent to join a terrorist group, or recruiting on behalf of a terrorist organisation • Graphically violent profile photos, Space icons, and cover photos. Examples include glorifying a violent mass casualty event (e.g., "I'm so glad [insert terrorist group] bombed that building"). Quora may also ban and remove all content from any user who is a confirmed and/or declared member of any group on the US State Department list offoreign terrorist organisations, or is a confirmed participant in acts of mass violence or hate crimes. **Hate speech** Hate speech is defined as a serious attack on a group or individual based on their race, ethnicity, gender, nationality, sexual orientation, sex, religion, caste, serious medical condition, or disability. This includes the use of slurs in a disparaging way, as well as content that dehumanises or calls for violence, exclusion, or segregation of protected classes.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.quora.com/about/tos and https://www.quora.com/about/acceptable_use
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No. There is no live-streaming functionality on Quora.
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 sole authority and final decision as to whether content or behaviour violates these policies. It may enforce these policies in a variety of ways, including, but not limited to written warnings, removal of content, adding warning tags to content, or the limitation or termination of a user's access to Quora.
4.1 Notifications of removals or other enforcement decisions	Users receive a notification from Quora Moderation if their content has been collapsed or deleted (Quora, 2021).
4.2 Appeal processes against removals or other enforcement decisions	If a user has received a notification from Quora Moderation about your content being collapsed or deleted, he or she can appeal the decision by clicking "Appeal" on the notification.

	To appeal other moderation decisions (such as edit-blocks or bans), users can use the "Contact Us" link located at the top of this page, and select "I want to appeal a moderation decision" from the drop-down (Quora, 2021).
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. Users have to use the 'report' option in the menu and select the option that best describes the policy violation. 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. In addition to human review, Quora uses automated processes for internal moderation. However, it does not provide further detail (Quora, n. d.). Quora is not a member of the GIFCT.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Content that violates Quora's Platform and Spaces polices 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). 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 private message and makes their case to be unblocked. Depending on the severity of the 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 account for violating any of its policies.
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).
12. Main changes since last Report	• Quora updated its Community Guidelines in 2022 and now explicitly prohibits acts of violence by terrorist organisations, whereas before it broadly prohibited 'glorifying or advocating violence'. The Guidelines feature new categories of violating content ('Harmful activities and self-harm') with more granular explanation and examples of what is allowed or not on the platform. • Users receive a notification if their content has been removed. • Users can appeal any content moderation decision, not just edit-blocks or bans.

28. 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 in its Section 3 on 'Code of Conduct' any activity that is harmful to the user, the service, or to others, such as posting terrorist or violent extremist content, communicating hate speech or advocating violence against others. It also prohibits users from publicly displaying or using Skype to share inappropriate content or material, involving for example graphic violence or criminal activity. Microsoft has stated (Microsoft, 2016) 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) that depicts graphic violence, encourages violent action, endorses a terrorist organisation or its acts, or encourages people to join such groups. The UN Sanctions List includes a list of groups that the UN 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
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 Skype), 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. 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 Skype, is brought to the company's attention via Microsoft's online reporting tool, Microsoft will remove it (Microsoft, 2016).
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 users have the opportunity to appeal account actions by visiting its appeals webpage (https://www.microsoft.com/en-us/DigitalSafety/moderation-and-enforcement/appeals) or using its appeals web from (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)	In addition, Microsoft has a dedicated reporting mechanism through which users can report 'terrorist content posted to a Microsoft consumer service'. Microsoft encourages users to use this form to report content posted by or in support of a terrorist organisation that depicts graphic violence, encourages violent action, endorses a terrorist organisation or its acts, or encourages people to join such groups (Microsoft, 2023). Microsoft deploys a variety of scanning technology, artificial intelligence, external partnerships, and human moderation operations solutions to detect and investigate TVEC. In particular, Microsoft explains that to detect potential child sexual exploitative content and/or terrorist and violent extremist content, it uses scanning technologies (e.g., PhotoDNA or MD5) and other AI-based technologies, such as text-based classifiers, image classifiers, and the grooming detection technique. 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). Microsoft is a founding member of the GIFCT and chairs the GIFCT Operating Board. Via the GIFCT, Microsoft participates in a range of activity, including engagement in its multi-stakeholder working groups and the GIFCT's Incident Response processes. In the event the GIFCT and its Operating Board activate a Content Incident or Content Incident Protocol, Microsoft ingests related hashes from the GIFCT's hash-sharing database. This allows Microsoft to quickly become aware of, assess, and address potential content circulating on its consumer services resulting from an offline terrorist or violent extremist event.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	See information in Section 4 above. 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)
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, 2020-2022). The latest transparency report covers the period from July to December 2022. 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. Microsoft also publishes an annual report under Regulation (EU) 2021/784 on addressing the dissemination of terrorist content online. The latest report is available on its CSR Trust Hub (https://www.microsoft.com/en-us/corporate-responsibility/reports-hub#coreui-feature-jy1t3q4) and Jurisdictional transparency reports page (https://www.microsoft.com/en-us/DigitalSafety/transparency-reports/jurisdictional-reports).
8. What information/fields of data are included in the TRs?	The latest transparency report (July - December 2022), includes the following information: • Number of TVEC actioned • Percentage of TVEC detected proactively • Number of accounts actioned due to TVEC • Percentage 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 actioned' refers to when Microsoft suspends or blocks access to an account, or restricts access to content within the account. 'Proactive detection' refers to Microsoft-initiated flagging of content on its services, whether through automated or manual review.

	'Accounts reinstated" refers to actioned accounts that were fully restored including content and account access, upon appeal.
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). Also, Europol highlighted that encrypted appliances, such as Skype, WhatsApp and Viber, offered relatively safe ways for terrorists to acquire illicit goods (weapons, fake IDs etc.) and services (Europol, 2016).
12. Main changes since last Report	Refreshed language on Microsoft's practices, processes and systems on addressing TVEC. Microsoft's also launched its new Digital Safety site (https://www.microsoft.com/en-us/DigitalSafety) with further details on its efforts to address TVEC.

29. Toutiao

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	There is no definition provided. However, Toutiao's User Agreement (Toutiao, s.d.) prohibits the production, copy, publication, or dissemination of content: • Promoting terrorism and extremism; • Promoting ethnic hatred and discrimination and undermining ethnic unity; • Spreading violence, murder, terror or abetting crimes; • Violently intimidating or threatening others; • Containing horror, violence and blood • That threatens life and health, uses knives and other dangerous instruments to endanger the personal and/or property rights of oneself or others • That encourages or induces others to participate in dangerous or illegal activities
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.toutiao.com/user_agreement/

3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No. Toutiao's User Agreement applies to all type of content uploaded on the platform, including live broadcast.
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?	Toutiao's User Agreement states that Toutiao has the right to independently judge and take advance warning, refuse to publish, immediately stop transmitting information, delete content or comments, ban the publication of content or comments for a short period, and restrict part or all of the account as appropriate. For suspected violations of laws and regulations or suspected illegal crimes, Toutiao will keep relevant records and has the right to report to the relevant authorities in accordance with the law, cooperate with the investigations of the relevant authorities, and report the case to the public security organs.
4.1 Notifications of removals or other enforcement decisions	Toutiao notifies users of enforcement decisions, via in-app notification or email.
4.2 Appeal processes against removals or other enforcement decisions	Users can appeal enforcement decisions made by Toutiao regarding content they have published.
5. Means of identifying TVEC (for example, monitoring algorithms, user-generated, human (staff) reviewers, hash-sharing/URL sharing database)	Toutiao uses a combination of human reviewers and automated tools to analyse posts and comments (Knight, 2017). 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 L. , 2019). Toutiao is not a member of the GIFCT.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	See section 4 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.
11. Has this service been used to post TVEC?	Unknown.
12. Main changes since last Report	Toutiao was not included in previous Reports.

30. 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 prohibits the use of its services to distribute content that promotes bigotry, racism, misogyny, and religious or ethnic hatred. Threatening others with violence is also prohibited. IMO's Community Guidelines (IMO, 2023) has explicit prohibitions of terrorism, violent extremism and hateful speech. Previously, the Guidelines featured a joint category for 'Terrorism and violent extremism' but they were updated in May 2023, and now the categories relevant to TVEC are as follows: • **Politics / Religion / Terrorism:** Under this policy, users are not allowed to produce or distribute any media that endangers local public security and/or falls afoul of laws and/or or promotes terrorism, including but not limited to: endangering national security and harming national interests; inciting national or racial enmity, hatred or discrimination; racial discrimination; insulting, spoofing or attacking religious policies/images or otherwise blaspheme religion, inciting religious opposition, disrupting social order, participating in illegal assemblies, and violating morality; spreading or propagandise terrorist organisation, terrorism or extremism tendencies, statements, photographs of terrorist leaders, etc. Users' accounts may be temporarily or permanently suspended if cases are confirmed. • **Violent extremism / Crime organising:** Under this policy, users are not allowed to produce or distribute any media that promotes, disseminates or propagandises wars, intimidation, crime abetting, violence, or violent extremism, including but not limited to media content related to hostage-taking by extremists, bloody violence content, etc. Users may

	not use IMO to promote, organise, abet or advocate any criminal behaviour in any form, including but not limited to: violence, theft, fraud, etc. Relevant content will be removed and accounts may be temporarily or permanently suspended if cases are confirmed. • **Hateful speech:** Users may not attack anyone based on their race, ethnicity, national origin, gender, gender identity, sexual orientation, religious affiliation, disabilities, or diseases. Relevant content will be removed and accounts may be temporarily or permanently suspended if cases are confirmed.
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. 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, 2023). 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.
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 violent 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.
12. Main changes since last Report	• IMO updated its Community Guidelines in May 2023. The category 'terrorism and violent extremism' does not exist anymore. In the present Guidelines, TVEC falls under two distinct categories: 'Politics / Religion / Terrorism' and 'Violent extremism / Crime organising'.

31. Xiaohongshu

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. Xiaohongshu's User Service Agreement prohibits TVEC. Notably, it does not allow: • Content endangering national unity, sovereignty and territorial integrity, leaking state secrets, endangering national security, damaging national dignity, honor and interests, promoting terrorism and extremism; • Content defaming the excellent cultural traditions of the nation, inciting ethnic hatred and discrimination, infringing on ethnic customs and habits, distorting

	national history and national historical figures, hurting national sentiments, and undermining national unity; • Inciting and undermining national religious policies, promoting religious fanaticism, endangering religious harmony, hurting the religious sentiments of religious citizens, undermining the unity between religious and non-religious citizens, and promoting cults and superstitions; • Endangering social morality, disrupting social order, undermining social stability, promoting obscenity, gambling, drug abuse, exaggerating violence and terror, instigating crimes or teaching criminal methods. In addition, Xiaohongshu's Community Standards (Xiaohongshu, 2021) prohibit the publication and dissemination of content that: • Spreads violence, murder, terror or instigates crime • Promotes terrorism or extremism or incites the implementation of terrorist or extremist activities • Incites ethnic hatred, ethnic discrimination, and undermines ethnic unity
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://agree.xiaohongshu.com/h5/terms/ZXXY20220331001/-1, and https://agree.xiaohongshu.com/h5/terms/ZXXY20221213003/-1
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	Yes. Xiaohongshu has separate Community Live-Streaming Standards for its live broadcast function (Xiaohongshu, 2023). Under these Standards, it is strictly prohibited to spread content that may endanger the life, health and property safety of users, including but not limited to: • Display and disseminate text, images, audio and video content involving violence, terror, and bloodshed etc. • Display and promote content that endangers the life, health and property safety of others, such as fighting, personal attacks, threats, bullying and other violent behaviours. The Guidelines also prohibit the dissemination and performance of behaviours that are likely to cause imitation by the audience and endanger life and health, including but not limited to:

	• Performances using high-risk items such as knives, simulated knives, guns, and simulated guns. Also, the Guidelines strictly prohibit to spread content that intentionally or may infringe the legitimate rights and interests of others, including but not limited to: • Inciting human flesh searches, deliberately guiding the exposure of other people's private information, provoking war attacks and other online violent behaviours. If a live broadcaster commits such violations, the platform will, based on the violation situation and the number of violations, issue a warning to the live broadcaster's account and/or the corresponding live broadcast room, interrupt the live broadcast, limit the live broadcast function, close the live broadcast permission, ban the account, or take other necessary measures. If the user does not agree, he or she can file an appeal. Users can use the report mechanism to report live broadcasters who violate Xiaohongshu's policies.
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?	In case of violation, Xiaohongshu has the right to make an independent judgement and immediately suspend or terminate the provision of some or all services, including banning speech, blocking information, deleting published content, banning accounts, cancelling accounts, and other measures.
4.1 Notifications of removals or other enforcement decisions	In case of violation, Xiaohongshu will notify users accordingly.
4.2 Appeal processes against removals or other enforcement decisions	If the user disagrees with an enforcement action, he or she can submit an appeal.
5. Means of identifying TVEC (for example, monitoring algorithms, user-generated, human (staff) reviewers, hash-sharing/URL sharing database)	Xiaohongshu encourages users to proactively report any content that may violate its policies. Xiaohongshu uses a combination of automated content moderation tools (Wanging, 2022) and human 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 L. , 2019). Xiaohongshu is not a member of the GIFCT.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	See sections 3 and 4 above. In 2022, a leaked document revealed how Xiaohongshu deals with censoring discourse about 'sudden incidents' on its platform. The document lists a range of particularly sensitive content that requires special treatment, including terrorist attacks, mass incidents, violent criminal cases targeted at specific populations such as fatal stabbings, but also demonstrations, student strikes, or widespread public criticism of government institutions. All high-risk incidents must be reported to community managers and the Shanghai Operation Security Group within five minutes (Boyd, How Xiaohongshu censors "sudden incidents", 2022).
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. Main changes since last Report	Xiaohongshu was not included in previous Reports.

32. Microsoft Teams

The Microsoft Services Agreement applies only to consumer use of Teams, and not to enterprise use. In an enterprise context, Microsoft acts as a data processor and the enterprise customer controls all customer content, including end user content – any rights for the service provider to access and/or process the organisational customer's content are defined in (and constrained by) the legal agreement.

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, Microsoft's Services Agreement, which governs Teams, prohibits any activity that is harmful to the user, the service, or to others, such as posting terrorist or violent extremist content, communicating hate speech or advocating violence against others. Microsoft has stated (Microsoft, 2016) that, for the purposes of its services, terrorist content is material posted by or in support of organisations included on the Consolidated United Nations Security Council Sanctions List (United Nations Security Council) that depicts graphic violence, encourages violent action, endorses a terrorist organisation or its acts, or encourages people to join such groups. The UN Sanctions List includes a list of groups that the UN 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
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. 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 Teams, is brought to the company's attention via Microsoft's online reporting tool, Microsoft will remove it (Microsoft, 2016).

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 users have the opportunity to appeal account actions by visiting its appeals webpage (https://www.microsoft.com/en-us/DigitalSafety/moderation-and-enforcement/appeals) or using its appeals web from (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)	Teams has a general reporting mechanism for users to report abusive comments or content (Microsoft, n.d.) In addition, Microsoft has a dedicated reporting mechanism through which users can report 'terrorist content posted to a Microsoft consumer service'. Microsoft encourages users to use this form to report content posted by or in support of a terrorist organisation that depicts graphic violence, encourages violent action, endorses a terrorist organisation or its acts, or encourages people to join such groups (Microsoft, 2023). Microsoft deploys a variety of scanning technology, artificial intelligence, external partnerships, and human moderation operations solutions to detect and investigate TVEC. In particular, Microsoft explains that to detect potential child sexual exploitative content and/or terrorist and violent extremist content, it uses scanning technologies (e.g., PhotoDNA or MD5) and other AI-based technologies, such as text-based classifiers, image classifiers, and the grooming detection technique. 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). Microsoft is a founding member of the GIFCT and chairs the GIFCT Operating Board. Via the GIFCT, Microsoft

	participates in a range of activity, including engagement in its multi-stakeholder working groups and the GIFCT's Incident Response processes. In the event the GIFCT and its Operating Board activate a Content Incident or Content Incident Protocol, Microsoft ingests related hashes from the GIFCT's hash-sharing database. This allows Microsoft to quickly become aware of, assess, and address potential content circulating on its consumer services resulting from an offline terrorist or violent extremist event.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	See information in Section 4 above. 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)
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, 2020-2022). The latest transparency report covers the period from July to December 2022. 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. Microsoft also publishes an annual report under Regulation (EU) 2021/784 on addressing the dissemination of terrorist content online. The latest report is available on its CSR Trust Hub (https://www.microsoft.com/en-us/corporate-responsibility/reports-hub#coreui-feature-jy1t3q4) and Jurisdictional transparency reports page (https://www.microsoft.com/en-us/DigitalSafety/transparency-reports/jurisdictional-reports).
8. What information/fields of data are included in the TRs?	The latest transparency report (July - December 2022), includes the following information: • Number of TVEC content actioned • Percentage of TVEC content detected proactively

	• Number of accounts actioned due to TVEC • Percentage 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 actioned' refers to when Microsoft suspends or blocks access to an account, or restricts access to content within the account. 'Proactive detection' refers to Microsoft-initiated flagging of content on its services, whether through automated or manual review. 'Accounts reinstated" refers to actioned accounts that were fully restored including content and account access, upon appeal.
10. Frequency/timing with which TRs are issued	On a semi-annual basis.
11. Has this service been used to post TVEC?	Unknown.
12. Main changes since last Report	Refreshed language on Microsoft's practices, processes and systems on addressing TVEC. Microsoft's also launched its new Digital Safety site (https://www.microsoft.com/en-us/DigitalSafety) with further details on its efforts to address TVEC.

33. 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. Viber's Acceptable Use Policy states, under the section titled 'Loving Vibes Prevail', that Viber has no tolerance for violent extremism or terrorism content and prohibits terrorist organisations or hate groups to use Viber, post content, sell or buy any goods, or communicate through the app. Viber recalls that it defends the right to express unpopular points of views, but does not allow hate speech which attacks or demeans a particular individual or group of people based on their ethnic or national origin, race, religion, disability, gender, gender

	identity, age or sexual orientation or other social or cultural factors. Viber adds that hate speech includes public incitement to violence or hatred directed against such people and can be provided orally, in text form, pictures or other media form. Hate speech can include publicly condoning, denying or grossly trivialising crimes of genocide, crimes against humanity and war crimes, when the conduct is carried out in a manner likely to incite violence or hatred against such a group or a member of such a group. Viber encourages users to share newsworthy content, content that relates to a matter of political, social, or other general concern to the Viber's community. However, Viber asks users to make sure the content is accurate, appropriate and factual. Content that promotes extreme political views, potentially used in the radicalisation of vulnerable members of the community is prohibited. Viber also strictly prohibits content that is related to terrorism, including the planning of a terrorist attack, promoting terrorist groups. Viber may remove such content, disable accounts and work with law enforcement agencies (as necessary under applicable law) when it believes that there is a genuine risk of physical harm or a direct threat to public safety under such circumstances. Moreover, under the section entitled 'Keep the Peace', Viber forbids users to encourage violence, use Viber to sell or buy any type of weapon, threaten to hurt a person or property, post any type of violent content, or encourage self-harm. Overly graphic expressions of violence, in any form including video clips, from games or films, in particular where the violence is glorified or encouraged, are not allowed. This includes extreme depictions or descriptions of violence, whether real or simulated, criminal activity, accidents and credible threats of violence or physical harm to any individual or group (Viber, 2022).
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/en/terms/viber-public-content-policy/
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	Not applicable. Viber does not have a live-streaming 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	Viber may remove any or all content if they deem that such content is unauthorized or illegal or violates Viber's Terms of Service. Viber also has an appeal process as described in Section 4.2 below.

enforcement decisions and appeal processes against them?	
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 Use Policy, in the event that Viber chooses to take action against any particular user with respect to any content that he or she has posted or that Viber decides to remove or refuse to distribute such content, the user may appeal or contest the decision to remove content or disable, block or suspend the user's account by contacting Viber through Viber Contact Us Form available at https://help.viber.com/en/contact. The user should include a reasoning as to why he or she feel that Viber's decision was incorrect. If Viber feels that its decision was in fact incorrect, it will notify the user of such and rectify the situation by putting the content back, reactivating the account or the Services for the user (as applicable) and removing any strikes or restrictions so that this will not be held against the user in the future."
5. Means of identifying TVEC (for example, monitoring algorithms, user-generated, human (staff) reviewers, hash-sharing/URL sharing database)	Viber has a user report system. Viber urges users and applicable authorities to report any content or activity which do not comply with its Terms of Service. It may also report directly to the authorities in certain cases and encourage users to do the same if the content can lead to a criminal event. Viber then reviews those reports and operates a moderation team to determine the most suitable course of action. In addition, Viber relies on administrators (or 'Admins') and superadmins to moderate and remove violating content from Communities and Channels: • Superadmins: Upon creating a Community or a Channel, a user automatically becomes a "Superadmin" of that Community or Channel. Superadmins can add, remove, or ban other Superadmins and Admins, or Members from the Community or Channel. They can also control who can post, choose whether the Community or Channel can be shared by other Members or not, pin or unpin messages, and delete messages by other Superadmins, They can also revoke invite links and enlist other Viber users to the Community or Channel to become Superadmins or Admins. • Admins: Although they have less control of the Community settings than a Superadmin, Admins play an important role in moderating and creating content for the Community. Some of their default permissions include the ability to ban members, delete messages, and pin

	messages to the Community chat. In channels admins play less of a moderation role, instead focusing more on content creation. Admins can post in the chat, pin messages, and be contacted by Channel members. There can be multiple Community or Channel Superadmins and Admins, with a total limitation of 250 combined. They must ensure that all content uploaded and displayed in Communities or Channels 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. Only the Superadmins and Admins of a Channel are be able to post content within such Channel – except in comment posts, if enabled, where members can add their comment on the specific message. If a member acts inappropriately, they have the right to remove or ban them from their Community. (Viber, 2023) Lastly, Viber further reserves the right to use artificial intelligence and machine learning tools as well as human teams to pre-moderate and review content published for any potential illegal content or content that violates its Terms of Service. (Viber, 2022) Viber is not a member of the GIFCT.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	According to Viber's Acceptable Use Policy, it may remove the offending content, terminate or limit the visibility of a user's account, or notify law enforcement. Viber may remove any reported content, at its sole discretion, if it finds it to be in breach of its Acceptable Use Policy, Terms of Service, or applicable law. Also, Viber further reserves the right to block or suspend a user from using its Services or the reporting feature if it determines that such user frequently submits reports or complaints that Viber has determined are manifestly unfounded.
7. Does the service issue transparency reports (TRs) on TVEC?	No.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.

10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Yes. ISIS announced (Site Intelligence Group Enterprise, 2018) a Nashir News Agency (the ISIS-linked media dissemination group) account on Viber (Katz, A Growing Frontier for Terrorist Groups: Unsuspecting Chat Apps, 2019). Viber closed the account immediately after finding it. In the case of the thwarted terrorist attack against two churches in November 2020 in Villejuif, France, the offender was using Viber to communicate on his smartphone (Le Journal du Dimanche, 2015).
12. Main changes since last Report	• Viber updated its 'Acceptable Use Policy' in March 2022 with more extensive and precise explanations of what constitutes terrorist, violent extremist, and hate speech content. In addition to prohibiting content that is related to terrorism, including the planning of a terrorist attack, promoting terrorist groups, it also prohibits terrorist organisations or hate groups to use Viber, post content, sell or buy any goods, or communicate through the app. • Pursuant to EU Code of Conduct against Hate Speech, Viber has also created a special page and procedure for Trusted Flaggers, according to which they can report hate speech in priority regime.

34. 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, under Section 4.3 on 'Prohibited Content', 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. Under Section 4.4 on 'Prohibited Conduct', Youku Tudou also prohibits to participate in any illegal or potentially illegal activities or transactions, including teaching criminal methods, selling illegal drugs, money laundering, fraud etc.
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?	Section 4 of Youku Tudou's ToS indicates that the platform has the right to manage the information that users upload, publish, or transmit. Youku Tudou will immediately stop transmitting prohibited content and take measures to prevent its dissemination, keep relevant records, and report to the relevant authorities.
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 L. , 2019). Youku Tudou is not a member of the GIFCT.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Section 8 on 'User's breach of contract' states that the platform has the right to immediately delete or block the corresponding content and information in accordance with the corresponding rule or suspend, seize, or cancel an account. Besides, if a user's behaviour constitutes a breach of contract, Youku Tudou may suspend the provision of some or all services to the user. In the most severe cases, Youku Tudou may permanently close the account and terminate the provision of services 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.
12. Main changes since last Report.	No main changes since last Report.

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 Guidelines, under the section titled 'Terrorism and Violent Extremism', 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. For example, users may not display or link terrorist or extremist violence, even for the purposes of denouncing such content. Moreover, the section titled 'Violence and Threats' states that violence is taken seriously on Twitch and is considered a zero-tolerance violation. All accounts associated with such activities will be indefinitely suspended. For examples, users may not show or promote: • Attempts or threats to physically harm or kill others • Attempts or threats to hack, dox, DDOS, or SWAT others • Use of weapons to physically threaten, intimidate, harm, or kill others • Encouraging others to participate in acts that may harm others Under the section titled "Hateful Conduct", Twitch does not permit behaviour that is motivated by hatred, prejudice or intolerance, including behaviour that promotes or encourages discrimination, denigration, harassment, or violence based on the following protected characteristics: race, ethnicity, colour, caste, national origin, immigration

	status, religion, sex, gender, gender identity, sexual orientation, disability, serious medical condition, and veteran status. Twitch also provides certain protections for age. For example, users may not: • Promote, glorify, threaten, or advocate violence, physical harm, or death against individual(s) or groups on the basis of a protected characteristic, including age. • Use hateful slurs, either untargeted or directed towards another individual. • Post, upload, or otherwise share hateful images or symbols, including symbols of established hate groups and Nazi-related imagery. • Create speech, imagery, or emote combinations that dehumanise or perpetuate negative stereotypes and/or memes. • Create content that expresses inferiority based on a protected characteristic, for example, statements related to physical, mental, and moral deficiencies. • Call for subjugation, segregation or exclusion, including political, economic, and social exclusion/segregation, based on a protected characteristic, including age. • Encourage or support 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. • Mock the event/victims or deny the occurrence of well-documented hate crimes, or deny the existence of documented acts of mass murder/genocide against a protected group. • Make 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.

	• Encourage the use of or generally endorsing sexual orientation conversion therapy.
	
	• Support, promote, or be a member of a hate group, including sharing hate group propaganda materials.
	
	• Create accounts dedicated to hate, such as through abusive usernames.
	
	Lastly, in 2022, Twitch introduced a new Username Policy (Twitch, 2022) which prohibits the following:
	
	• References to terrorism or terrorist organisations
	• Threats, promotions, or calls to real-life violence against others (with exceptions for references to video game or non-hateful historical violence)
	  ○ Calling for another group of people to be harmed or killed
	  ○ Creating a username that threatens violence against another person
	• Hateful Conduct, including slurs and derogatory terminology related to protected characteristics
	  ○ Creating a username that includes a hateful slur
	  ○ Glorifying, promoting, or advocating for discrimination, denigration, segregation, exclusion, hatred, or disgust based on protected characteristics
	  ○ Creating a username that references a hate group
	  ○ Mocking, denying, or glorifying the occurrence of well-documented hate crimes or acts of genocide

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	Twitch takes enforcement action against accounts that violate its ToS and/or Community Guidelines. Twitch considers several factors when reviewing reports of

there notifications of removals or other enforcement decisions and appeal processes against them?	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 offences, an indefinite suspension. A warning is a courtesy notice for some violations. 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 one to 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 offences, Twitch immediately and indefinitely suspends the account with no opportunity to appeal. Twitch also notes that in exceptional circumstances, it 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 offences committed by members of the Twitch community (or those wishing to join it) 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, issues enforcements against the relevant users. Twitch may bring in a third-party investigator for impartial review. If Twitch is able to verify reports of off-service statements or behaviours that relate to an incident that took place on the platform, it will use this evidence to inform enforcement decisions (Twitch, 2023).

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. Appeals can be requested for enforcements issued within the last 60 days, or for a user's most recent enforcement if he or she is serving an indefinite suspension. In the appeal, the user must include the reason they believe the decision was incorrect. Once the appeal has been reviewed. For suspensions of 30 days or less, users may only submit one appeal per enforcement. For indefinite suspensions, users may only submit one appeal in a 6-month period. Twitch reviews appeals in the order they are received and does not guarantee enforcements will be overturned. Abusing or spamming the appeals process may lead to additional penalties or revoking of appeal privileges.
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 content and behaviour are allowed or not on Twitch. Then there is service-level safety, composed of three parts: machine detection, user reporting and review and enforcement: - Machine Detection: Twitch has implemented 'machine detection' technologies that scan content on the service and then flag it for review by human specialists (this is also called 'proactive detection'). 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 detection 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 enables 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 (or 'mods'), who moderate chat in the creator's channel (moderators can be easily identified in chat by the green sword icon that appears next to their username). Mods play many roles, from welcoming new viewers to the channel, to answering questions, to modelling and enforcing channel-level standards. Twitch provides both creators and their moderators with a powerful suite of tools such as AutoMod, Chat Modes, account verification, Suspicious User Detection, and Mod View to make their roles as easy and intuitive as possible. These tools provide the ability to manage who can participate in chat, automatically filter unwanted messages before they are displayed, give users "timeouts" (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. Lastly, there is the viewer-level safety, where viewers are able to customise the safety of their experience. Twitch provides features – such as mature flags, chat filters, and blocking other users – that can be used to customise content and interactions across the service. (Twitch, 2023) 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 offences, an indefinite suspension from Twitch. In addition, usernames that are flagged as violating the Community Guidelines, and do not result in an indefinite suspension, are forced to rename. If any content that contains the violation has been recorded on the service, Twitch will remove it. All suspensions are binding until expiration or removal upon successful appeal. Any attempt to circumvent an account suspension or chat ban by using other accounts, identities, or by appearing on another user's account will also result in an additional enforcement against the suspended user, up to an indefinite suspension.
7. Does the service issue transparency reports (TRs) on TVEC?	Yes (Twitch, 2022). The latest available TR covers H2 2022.
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 members of the Twitch ecosystem (streamers, moderators, viewers, and Twitch's staff). Then, Twitch issues "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:
	
	Moderation in channels: coverage, removals and enforcements
	
	• Moderation of chat: Percentage of minutes watched moderated, broken down per method of moderation (AutoMod, Moderation, Moderation or AutoMod). The report shows that Twitch is increasingly relying on automated tools (AutoMod) for content moderation. This increase is likely largely attributable to a change in the default settings for new channels, which now sets AutoMod to Level 1 if the channel has a moderator added, and to Level 2 if it does not.
	• Proactive and manual removals of chat messages:
	  o Number of removed messages, broken down per method of detection (manual or proactive)
	  o Number of deletions per 1000 messages
	• Channel enforcement actions:
	  o Number of actions taken, broken down per types of action (timeouts or channel bans)
	  o Number of enforcement actions per channel
	
	Reports made on Twitch
	
	• User reports
	  o Number of user reports broken down per category of violating content, among which 'Terrorism, Terrorist Propaganda, and Recruitment', 'Extreme Violence, Gore, and Other Obscene Conduct', and 'Hateful Conduct
	  o Number of user reports per 1000 hours watched broken down by types of violating content
	  o Response times: percentage of reports per response time (under 10 min, 30 min, 1 hour, 6 hours, 12 hours or 24 hours). This is not broken down per category of violating content.
	• Twitch launched in May 2022 a new report process which makes it easier for viewers to find the report reason they were looking for, which likely impacted the distribution of reports received as users could now more accurately classify their reports.

	Enforcements • Total number of enforcement actions • Total number of enforcement actions per 1000 hours watched • Number of enforcement actions per category of violating content, including 'Terrorism, Terrorist Propaganda, and Recruitment', 'Extreme Violence, Gore, and Other Obscene Conduct', and 'Hateful Conduct, Sexual Harassment, and Harassment • Number of enforcement actions per 1000 hours watched, per category of violating content, including 'Terrorism, Terrorist Propaganda, and Recruitment', 'Extreme Violence, Gore, and Other Obscene Conduct', and 'Hateful Conduct, Sexual Harassment, and Harassment' **User appeals** • Number and percentage of incoming appeals broken down per category of violating content, including 'Terrorism', 'Violent Graphic Content', and 'Hate/Harassment' • Number of appeals granted for each category • Twitch explains that the successful appeals for terrorism-related content were based on enforcement errors for news or documentary footage that referenced but did not depict, glorify, encourage, or support terrorism or violent extremist acts or actors. **Law enforcement and government requests** • Number of NCMEC CyberTips sent • Number of NCMEC CyberTips sent per 1000 hours watched • Number of escalations to law enforcement • Number of escalations to law enforcement per 1000 hours watched • Number of subpoenas & preservation holds processed • Number of subpoenas & preservation holds per 1000 hours watched processed Besides, in 2022, Twitch conducted a Human Rights Impact Assessment available at: https://safety.twitch.tv/s/article/Twitch-HRIA-2022?language=en_US

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 a semi-annual basis.
11. Has this service been used to post TVEC?	Yes. See section 8 above. On 9 October 2019, a far-right neo-Nazi terrorist killed two people outside a synagogue in the German city of Halle, and livestreamed on Twitch a 35-minute video of the attack that was seen by about 2,200 people before it was removed from the platform (Noack, Beck, & Morris, 2019). Moreover, on 14 May 2022, a white supremacist opened fire in a Buffalo, New York, supermarket. Part of the terrorist attack was livestreamed on Twitch, but it was shut down by the service in under two minutes, after being seen by 22 people. However, the video was reposted on other platforms, like Facebook and Streamable, and was seen by millions before being completely removed. As a result of the terrorist attack, ten people, all of whom were Black, were murdered and three were injured (Grayson, 2022).
12. Main changes since last Report	• In 2022, Twitch introduced a new Username Policy which prohibits references to terrorism or terrorist organisations, threats, and hateful conduct, in account usernames and display names. It also launched a new machine learning model to catch violative usernames and automatically enforce against them or prompt them to reset, depending on the severity of the violation. • Twitch launched in May 2022 a new report process which makes it easier for viewers to find the report reason they were looking for, which likely impacted the distribution of reports received as users could now more accurately classify their reports. • A list of Twitch's updates since the last transparency report is available in its H2 2022 transparency report: o Twitch updated its Community Guidelines with simplified language, more clarity, and additional examples. o Twitch now includes a "Shared Ban Info" tool, allowing streamers to share information about who they've banned in their channels with one another, to help

	keeping serial harassers out of their communities. ○ Twitch now features a Shield Mode that enables streamers and theirs mods to pre-set multiple safety settings that can be activated with a single click whenever they feel they need a higher level of protection. While Shield Mode is activated, specific terms and phrases can be bulk banned so that harassing messages and users can be easily removed from the chat. • Twitch's transparency report includes the following new metrics: ○ Number and percentage of incoming appeals broken down per category of violating content, including 'Terrorism', 'Violent Graphic Content', and 'Hate/Harassment' ○ Number of appeals granted for each category ○ Response times: percentage of reports per response time (under 10 min, 30 min, 1 hour, 6 hours, 12 hours or 24 hours). This is not broken down per category of violating content. • Twitch's latest transparency report shows that Twitch is increasingly relying on automated tools (AutoMod) for content moderation. This increase is likely largely attributable to a change in the default settings for new channels. • Twitch now issues transparency reports on a semi-annual basis (and not on an annual basis as previously). • In March 2022, Twitch launched a new appeals portal, which provides visibility into the enforcements that are eligible for appeal, and also displays the status and outcome of ongoing and prior requests, respectively.

36. ShareChat

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, ShareChat's Content Policy ("Community Standards/Guidelines") (ShareChat, 2021) explicitly prohibits terrorism and violence including organised violence. ShareChat's policies are similar to Moj's. **Violence:**

	ShareChat states that it prohibits violence, which includes all content that causes discomfort to users due to the goriness in the content, such as but not limited to graphical images or videos that glorify violence and suffering, or intends to incite violence, depiction of physical violence or animal cruelty. Content which promotes dangerous and illegal activities, or praises individuals, groups or leaders involved in terrorism, organised violence or criminal activities is strictly prohibited. Educative or informative content pertaining to violence may be allowed on the platform. Violent content on the platform in the form of fictional set up or martial arts may be permitted. **Illegal Activities:** ShareChat has zero-tolerance for content that advocates or promotes illegal activities. It prohibits content related to organised crime, criminal activities, promotion/sale/use of weapons, firearms and explosives, violence or terrorist activities. Users are not allowed to post content that displays tutorials or instructions or educates the users about illegal and prohibited activities including, but not limited to participating in criminal activities, making bombs or encouraging or doing or trading in drugs. Users must not use the platform to solicit or facilitate any transaction or gift involving such goods and services which are declared illegal by the Government of India. **Hate Speech and Propaganda:** Content that promotes violent behaviour against an individual or a group of individuals, intends to intimidate, target or demean any particular religion, race, caste, ethnicity, community, nationality, disability (physical or mental), diseases or gender, is prohibited. Any kind of content which produces hatred or has the intention of creating or spreading hatred or hate propaganda along the lines of including, but not limited to religion, caste, ethnicity, community, sexual orientation, or gender identity is also not allowed. ShareChat does not entertain content that spreads discrimination, intends to justify violence based on the above-mentioned attributes and refers to an individual or a group of individuals as inferior in any sense or with negative connotations. ShareChat urges users to refrain from incendiary commentary and publishing theories or hateful ideologies that may cause outrage to other users and influence them negatively. ShareChat may permit such content which intends to raise awareness about these issues or challenge it,

	subject to clear intention of posting such content on the platform. Lastly, ShareChat's Terms of Use (ShareChat, 2021) state that users shall not use the platform to share any content which is obscene, pornographic, harmful for minors, discriminatory, spreading what may be considered as hate speech, inciting any form of violence or hatred against any persons, or violates any laws of India, or is barred from being shared by any laws of India. ShareChat reserves the right to remove such content.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://help.sharechat.com/policies/content-policy/ and https://help.sharechat.com/policies/terms/
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	Yes. ShareChat's Terms of Use state that all content uploaded using the Lives feature is subject to the ShareChat Content and Community Guidelines. It reserves the right to immediately remove any content uploaded using this feature which is in violation of its Terms of Service including the ShareChat Content and Community Guidelines. ShareChat does not warranty the continued availability of this feature.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	ShareChat actively removes content which is not allowed and violates both its Guidelines as well as applicable Indian laws. If such content comes to ShareChat's attention, it may take it down or ban user accounts. ShareChat also encourages users to report content that may be in violation of its Community Standards. ShareChat adds that the intent of the creator is important. While it understands the importance of creative freedom, it does not welcome content that intends to bring discomfort, spread what may be considered hate speech and abuse or promote violence and illegal activities.
4.1 Notifications of removals or other enforcement decisions	ShareChat notifies users when their content is removed and specifies the reason for removal.
4.2 Appeal processes against removals or other enforcement decisions	If users believe that their content has been unfairly removed, they may raise an in-app appeal via the 'Violations' page accessible through the Profile Settings option on the users' profile. They may also contact grievance@sharechat.co to dispute the action. ShareChat may reassess the content and determine the validity of appeals raised by the user.
5. Means of identifying TVEC (for example, monitoring algorithms, user-	ShareChat uses a combination of human moderators and deep learning-based Computer Vision models (commonly known as AI) to implement its Community Standards. These AI model(s) facilitate the detection of harmful content on

generated, human (staff) reviewers, hash-sharing/URL sharing database)	ShareChat. Once flagged by the model, such content goes through human moderators, who decide the action to be taken as per various standards set forth by the Guidelines and standards assigned to the model. ShareChat's approach to content moderation comprises the following: • Proactive approach: Each post generated on ShareChat undergoes scrutiny by a series of AI model(s) to calculate a risk score. These models typically flag content based on assigned labels such as sexually explicit material, violence, or the presence of blood/weapons. Following the risk scores assigned to these predictions, content with high-risk scores is prioritised for manual review. Subsequent actions taken on flagged content and the priority for manual review are determined by the severity of the violation. For instance, 'Not Safe For Work' (NSFW) content may be prioritised higher in manual moderation and may invoke a stricter ban on the user. This manual review of flagged content not only aids in reducing false rejections but also contributes to fine tuning the model, as automated actions may be taken regarding similar posts that are in violation of the Community Standards. • User feedback: Apart from the above-mentioned flagging system, content that is in violation of ShareChat's Community Standards can also be reported by logged in and non-logged in users. Users can report both user profiles and individual content using the in-app reporting mechanism across multiple categories including violence, illegal activities, terrorism, hate speech, and more. In addition to the in-app reporting mechanism, users can also raise reports through other reporting mechanisms as mentioned below: o In-app Help and Support Chat: Registered/ logged-in users on ShareChat can go to their profiles and access the Help and Support Chat and report any violation of the Community Standards; o Email Support: An individual can also contact the ShareChat support team via email. The relevant email id is contact@sharechat.co; o Grievance Officer: ShareChat has a Grievance Officer to address platform usage

	concerns. The relevant email ID is grievance@sharechat.co. These reports are reviewed, and appropriate actions are taken by ShareChat's user grievance and content moderation teams, in accordance with the Community Standards and applicable rules and regulations. ShareChat also uses relevant content reports for bettering their AI model. • Reactive approach: Another strategy developed at ShareChat is to routinely review highly viral content as such content may impact a large number of users. Every post that crosses a preset threshold of virality is reviewed manually by the human moderators. Automated content moderation tools used by ShareChat include for example label propagation (i.e., propagating the labels from a few set of labelled examples to unlabelled examples based on their similarity with the labelled examples), active learning, semi-supervised learning, classifiers, vision models to process images and videos etc (Mandav, Parihar, Saket, Gupta, & Mukherjee, 2021). ShareChat is not a member of the GIFCT.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	ShareChat takes strict and prompt action against those who violate its Guidelines. If a profile is reported for violation, it may be suspended temporarily. In case of repeated breach of the Guidelines or attempt to circumvent a suspension, ShareChat may be compelled to terminate the account permanently and block the user from registering again. If required, ShareChat will cooperate with legal authorities and law enforcement mechanisms. ShareChat enforces different types of account-level bans that can range from hourly bans (3- and 6-hour bans) to one day, three days, seven days, 30 days, and 360 days: • User-generated content ban: User is unable to post any content on the platform for the specified ban duration. • Edit profile ban: user is unable to edit any profile attributes for the specified time period. • Comment ban: user is banned from commenting on any post on the platform for the specified ban duration.

	• Chatroom ban: User is unable to use the chatroom feature (ShareChat, 2023).
7. Does the service issue transparency reports (TRs) on TVEC?	Yes. ShareChat publishes monthly transparency reports. The latest available report covers the month of September 2023 (ShareChat, 2023).
8. What information/fields of data are included in the TRs?	ShareChat's TR includes one TVEC-specific information: • Number and percentage of content and accounts reported, based on various categories as tagged by users such as 'Terrorism', 'Hate speech', 'Illegal activities', and 'Violence'. The TR includes other non-TVEC-specific information: **Law enforcement requests:** • Law enforcement requests received • Requests where user data was provided • Requests wherein a takedown/ban action was taken on the basis of violation of the community guidelines **User reports:** • Total number of user reports received • Number and percentage of user reports (content and accounts) broken down by report reasons **Enforcement actions (takedown)** • Number of content pieces taken down by categories of content **Enforcement actions (profile bans)** • Number of user-generated content bans, edit profile bans, comment bans, chatroom bans, broken down by ban duration (hourly, 1 day, 3 days, 7 days, 30 days, 360 days), applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	No information is provided.
10. Frequency/timing with which TRs are issued	On a monthly basis.
11. Has this service been used to post TVEC?	Yes. See section 8 above.

12. Main changes since last Report	ShareChat was not included in previous Reports.

37. 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 excessively violent content, and expressions that lead to discrimination by race, antional origin, creed, gender, social status, family origin, etc. 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 states that any violating posts are suspended once detected. LINE also bars users who repeatedly breach the rules from using its services, or close their accounts.
4.1 Notifications of removals or other enforcement decisions	No information is provided.
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 all reviewed by LINE's patrol team and they 'take appropriate action' (LINE, n.d.) if they find any violations of such policies. The LINE Group has monitoring centres at five offices across four countries that constantly patrol our public services 24 hours a day, 365 days a year. These centres check every public post made in Japanese, English, Chinese, Thai, and Indonesian and promptly suspend any that are identified as violations. In addition to responding to the user reports, LINE uses a combination of AI and humans to monitor user-generated content and identify violations (LINE, 2022).

	First, user-posted content on supported LINE services is checked by an AI-driven system to automatically assess whether text, photos and other posts could be in violation of LINE's standards. If objectionable content is found by the monitoring system, it is immediately suspended after being posted. Next, a monitoring team manually 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, 2023). LINE does not monitor private content such as LINE chats and non-public VOOM (LINE's video platform) posts. However, users are able to report this type of content if they feel it is violating LINE's guidelines. Only in this case will the patrol team review the reported private content (LINE, 2022). In 2020, LINE became the first Asia-based company to join the Christchurch Call to Action (LINE, 2020). LINE is not a member of the GIFCT.
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 ('content moderation reports'), and message and call encryption deployment status (LINE, 2023). The last content moderation report (covering H1 2022) was issued in 2023, and contains no specific information on TVEC. Although it is not specified, TVEC could be included in the categories of 'illegal activities' and 'harmful/dangerous content'. It should also be noted that the reports now cover a total of 22 services (against 10 services in H1 2019).
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.

10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Unknown.
12. Main changes since last Report	• LINE updated the categories of violating content used in its transparency reports. In addition to the categories under which TVEC may fall that already existed ('disturbing and problematic content' and 'illegal activities'), it created a new separate category for 'harmful/dangerous content' which covers declarations to commit a crime (such as murder or bombing, risks to human life, or causing chaos among the general public. However, TVEC is still not specifically mentioned.

38. 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, spreading or disseminating violence, murder, terror or abetting crimes, propagating ethnic hatred, ethnic discrimination, and undermining ethnic unity, violently intimidating or threatening others. The ToS also prohibit content that contains horror, violence and blood, high risk, and endangers the physical and mental health of users themselves or others.
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?	The ToS apply to the live video service available on Xigua Video.
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?	Violation of Xigua's ToS may lead to the termination of the infringer's account and access to Xigua's services, 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	Xigua Video now has an appeal mechanism in place. If users do not agree with a decision made the platform regarding content they have published, they can appeal the 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 Xigua. While 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, I helped build ByteDance's vast censorship machine, 2021). Xigua is not a member of the GIFCT.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	In case of violation of the ToS, Xigua Video states that the company has the right to independently judge and take pre-warnings, refuse to post, immediately stop transmitting information, delete content or posts, prohibit short-term posting of content or posts, and restrict parts or all functions of the account until the termination of service provision, permanent account closure and other measures.
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. Main changes since last Report	• Xigua Video now has an appeal mechanism in place, available to users who do not agree with a decision made by the platform regarding content they have published. • Xigua Video provides more information on sanctions in case of violation of its ToS: it has the right to independently judge and take pre-warnings, refuse to post, immediately stop transmitting information, delete content or posts, prohibit short-term posting of

	content or posts, and restrict parts or all functions of the account until the termination of service provision, permanent account closure and other measures (whereas before it only mentioned termination of the infringer's account and access to Xigua's services).

39. 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's Terms of Service and Vimeo's Acceptable Use Community Guidelines prohibit any content that: - promotes or supports terror or hate groups; - any content that depicts unlawful real-word acts or extreme violence, vivid, realistic, or particularly graphic acts of violence and brutality; - provides instructions on how to assemble explosive/incendiary devices or homemade /improvised firearms; - contains hateful or discriminatory speech; - violates any applicable law. Vimeo also prohibits the use of its services by gangs, hate groups, terror organisations, members of the foregoing, and persons who are subject to US sanctions. In certain cases, persons who are subject to sanctions by a non-US government may also be included in this category. In addition, if users are in a country or region that is subject to comprehensive US sanctions, they may not purchase software services or hardware from Vimeo. Vimeo defines hate and terror groups as organisations that "aim to spread propaganda designed to radicalise and recruit people or aid and abet attacks". Vimeo defines hate groups as organisations that have adopted, based upon its statements, leaders or activities, a hateful ideology. More specifically, Vimeo defines a hateful ideology as a set of beliefs that malign a group based upon personal characteristics. For US-based groups, Vimeo considers the Southern Poverty Law Centre's designations of hate groups to be conclusive. For non-US 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 US Federal Bureau of Investigation's list of domestic terror groups and the US Department of State's list of foreign terror groups as conclusive, but not exhaustive. A gang means any organisation 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. Vimeo defines hateful and discriminatory speech as any expression that: - is directed to an individual or group of individuals based upon personal characteristics of that individual or group; - conveys a message of inferiority or contempt; and - would be considered extremely offensive to a reasonable person. Personal characteristics are core elements of identity that are shared by groups of people (and are generally not specific to any one person) and include: race, colour, national origin, and ethnicity, gender identity sexual orientation, religion, disability, and age. 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 Vimeo's definition covers, for example, videos that assert harmful stereotypes, claim racial superiority of one group over another, or suggest that certain groups of people of a particular religion are involved in far-flung conspiracies (Cheah, 2019).
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://vimeo.com/terms and https://vimeo.com/help/guidelines; and also in Vimeo's Help Center: A Comprehensive Guide: What content you can upload to Vimeo
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	The ToS apply to Vimeo's livestream services.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular, are there: notifications of removals or other enforcement decisions and appeal processes against them?	Vimeo states that context is of the essence in the application of its rules and processes. When prohibited content appears in the context of a news story or a narrative device in a dramatic work, Vimeo is likely to leave it up. If, however, the overall driving message of the work is to perpetuate a viewpoint that Vimeo has specifically banned, 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). 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, 2023).
4.1 Notifications of removals or other enforcement decisions	Enforcement decisions are notified via email except in cases of CSAM, terrorist content, fraud, spam, sextortion or other illegal content on 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, by either flagging it (where provided) or contacting Vimeo. Moderators review these reports and take action accordingly (Vimeo, 2023). Art 5.1 (Approach to Moderation) of Vimeo's Acceptable Use Community Guidelines explains that Vimeo endeavours to review specific content that is flagged by users, third parties, and certain software-based systems. Vimeo does not endeavour to review every piece of content uploaded to its systems. In addition, it explains that, when it does review content, it is usually for a particular reason, and so it does not endeavour to review it

	for all possible terms violations. Nor does it "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).
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Violations may result in suspension or removal of videos, account privileges, or the entire account. Account removal will occur in severe cases, such as where users have wilfully or repeatedly violated Vimeo's terms, or have uploaded extremely inappropriate content. If an account is permanently removed, the user may not create a new account. Vimeo may, where appropriate, grant grace periods to comply with its requirements. The failure to address the concerns within the provided timeframe will be considered a violation itself absent good cause.
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?	According to Vimeo, it has happened on rare occasions.
12. Main changes since last Report	No main changes since last report.

40. 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 (Discord, 2023) prohibit the use of Discord for the organisation, promotion or support of violent extremism. This also includes glorifying violent events or the perpetrators of violent acts, as well as promoting conspiracy theories that could encourage or incite violence against others. 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). Examples include racially motivated violent groups, religiously motivated groups dedicated to violence, and incel groups. Discord's Community Guidelines does not allow to use hate speech or engage in other hateful conduct. This includes the use of hate symbols and claims that deny the history of mass human atrocities. Discord considers hate speech to be any form of expression that either attacks other people or promotes hatred or violence against them based on their protected characteristics: age; caste; colour; disability; ethnicity; family responsibilities; gender; gender identity; housing status; national origin; race; refugee or immigration status; religious affiliation; serious illness; sex; sexual orientation; socio-economic class and status; source of income; status as a victim of domestic violence, sexual violence, or stalking; and weight and size. Lastly, threats to harm to another individual or group of people are prohibited. This includes direct, indirect, and suggestive threats.
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	Discord explains that violation of its Community Guidelines or other policies enables them to take a 'number of enforcement steps', including issuing warnings, removing content, suspending or removing the accounts and/or servers (equivalent to groups or communities under a

enforcement decisions and appeal processes against them?	common theme) responsible, and potentially reporting them to law enforcement. Since 2022, Discord may also consider relevant off-platform behaviour when assessing for violations of specific Community Guidelines (Discord, 2022).
4.1 Notifications of removals or other enforcement decisions	No notifications are specified.
4.2 Appeal processes against removals or other enforcement decisions	Users can appeal any enforcement action taken against their accounts, including terminations, suspensions, or content removals (https://dis.gd/request). 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)	Discord currently employs three levers to moderate user content: • User controls: Discord's product architecture provides each user with fundamental control over their experience including who they communicate with, what content they see, and what communities they join or create. • Platform moderation: Community Guidelines apply to all content and every interaction on the platform. These fundamental rules are enforced by Discord on an ongoing basis through a mix of proactive and reactive work. In Community Servers larger than 200 members, Discord takes a more proactive and automated approach to safety, and may use automated means to detect violations of its policies. • Server owners and volunteer community moderators define and enforce norms of behaviour for their communities that can go beyond the Community Guidelines. Discord enables community moderators with technology (tools like AutoMod) as well as training and peer support (Discord Admin Community). Users can report any content that violates Discord's ToS and Community 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 or user (Liao S. , 2018). After the report, Discord's 'Trust and Safety' team

	acts as detectives, looking through the available evidence and gathering as much information as possible. This investigation centres on the reported messages, but can expand if the evidence shows that there is a bigger violation – for example, if the entire server is dedicated to bad behaviour, or if the behaviour appears to extend historically. In 2021, Discord's 'Trust & Safety' team made up 15% of its near 400 employees. Since 2022, Discord takes the off-platform harmful behaviour into consideration when assessing whether that account has violated a specific Community Guideline. The Trust & Safety team may launch an investigation of an account, including reviewing the user's activity and posts, based on this off-platform behaviour. These measures apply only in the case where Discord becomes aware of highest-harm threats, including organising, promoting, or supporting violent extremism, making threats of violence, carrying out acts of violence, or sexualising children in any way. Harmful off-platform include, among others: • Recruiting or participating in activities within a known violent group • Explicit and/or credible threats of violence towards a person, group of people, organisation, event, or location • Carrying out acts of violence, such as a mass shooting or acts of human trafficking (Discord, 2022) With regard to violent extremism, 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'. While Discord might use these signs to determine a user or space's intent or purpose, it always wants to understand the context in which user content is posted before taking any action (Discord, 2021). Discord also relies on user-moderators who are in charge of the different communities. In 2021, Discord 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). Besides, Discord uses proactive tooling and resources to ensure that violent and hateful groups do not find a home on the platform. In 2021, Discord acquired an AI-based software company called Sentropy to expand its ability to detect and remove bad content, including hate, violence, and other forms of harm (Discord, 2021). In June 2022, Discord introduced AutoMod, a customisable moderation tool equipped with a keyword filter that can automatically detect, block, and alert users of messages containing harmful words or phrases before they are posted in text channels, threads, or Text Chat in Voice (Discord, 2022). Discord is a member of the GIFCT.
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 “cooldown” period - Permanently banning users from Discord and making it difficult for them to create another account - Removing a server from Discord (Discord, 2023) - 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 from October to December 2022, contain specific metrics on violent extremist content removal. This is Discord’s tenth transparency report since 2019 (Discord, 2023).
8. What information/fields of data are included in the TRs?	Discord’s last transparency report discloses: **Actions taken on accounts and servers** • Total number of warnings sent to individual accounts, to servers, and to server members, each of them broken down by category of violating content, among which ‘violent extremism’, ‘violent and graphic content’, and ‘hate speech’ • Total number of accounts disabled, broken down by category of violating content • Total number of servers removed, number of servers removed proactively v. reactively, broken down by category of violating content • Number and percentage of accounts who submitted an appeal, broken down by category of violating content • Number and percentage of appeals granted, broken down by category of violating content • Percentage of accounts unbanned • Number of legal requests from law enforcement, and number of requests complied • Number of emergency disclosure requests from law enforcement, and number requests complied **Reports** • Total number of user reports received • Number of user reports received, broken down by category of violating content, among which ‘violent

	extremism', 'violent and graphic content', and 'hate speech' • Number and percentage of reports actioned, broken down by category of violating content Discord's transparency report features a section titled 'Enforcement trend analysis' which gives an idea of the recent evolution regarding content moderation on the platform. For instance, in Q4 2022, Discord removed 7,223 accounts and 829 servers for violent extremism, a decrease of 42% and 21% respectively. The rate of proactively removing servers for violent extremism increase by 20% during this quarter, to reach 65%.
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 quarterly basis (instead of semi-annually as previously).
11. Has this service been used to post TVEC?	Yes. See Section 8 above. Most recently, on 14 May 2022, a gunman targeted a predominantly Black community in Buffalo, New York, and carried out a mass shotting in a supermarket. Before the attack, the perpetrator had written a white supremacist manifesto in his private Discord server – visible only to him. However, 30 minutes before the attack, he shared invitations to view his private server with a small number of other users and servers (Discord, 2022). In the weeks before the storming of the US Capitol on 6 January 2021, Discord's 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 the insurrection (Discord, 2021).
12. Main changes since last Report	• In 2022, the perpetrator of the mass shooting in Buffalo, New York, wrote a white supremacist manifesto on his Discord private server, which he shared with a small number of users right before the attack. • Discord provides additional information about its detection methods: o After a user 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. In 2021, Discord's 'Trust & Safety' team made up 15% of its near 400 employees. ○ In Community Servers larger than 200 members, Discord takes a more proactive and automated approach to safety, and may use automated means to detect violations of its policies. • In June 2022, Discord introduced AutoMod, a customisable moderation tool equipped with a keyword filter that can automatically detect, block, and alert users of messages containing harmful words or phrases before they are posted in text channels, threads, or Text Chat in Voice. • Since 2022, when it becomes aware of highest-harm threats (including organising, promoting, or supporting violent extremism, making threats of violence, carrying out acts of violence, or sexualising children), Discord takes the off-platform harmful behaviour into consideration when assessing whether that account has violated a specific Community Guideline. • Discord now issues transparent reports on a quarterly basis (instead of semi-annually).

41. Josh

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	Josh defines terrorist and extremist content as 'any content which threatens the security and national integrity of India or promotes any dangerous activities are a violation of Josh's Platform's Policies' (Josh, 2022). Josh's Community Guidelines prohibit 'Terrorism and Extremism'. Under this policy, users must not threaten or promote any dangerous activities on the platform. In particular, they must not use the platform to incite terrorism, secession, acts of violence against person or property or which threatens the unity, integrity, defence, security or sovereignty of India, friendly relations with foreign nations, or insults another nation. Users must not post any content that promotes or encourages users to take actions on behalf

	of a terrorist or anti-national organisation, or recruits and disseminates information for, or furthers the objectives of, such organisations. Users must also not garner support or approval for the commission of violent acts during lawful protests or create violence-inducing terror and conspiracy networks on the platform. In addition, Josh's Community Guidelines prohibit 'Hatespeech and Discrimination'. Users are required to trat others with dignity, respect and a sense of empathy. Josh does not allow any forms of hateful, personal attacks, ad hominem speech, or uncivil disagreement which are intended to harm another user or cause them mental stress or suffering. Examples of hateful or discriminatory speech include comments which encourage violence, racially or ethnically objectionable, or disparage anyone based on their national origin, sex, gender identity, sexual orientation, religious affiliation, disabilities, or diseases. Josh also prohibits content which could encourage other users to share such content. Users are not permitted to use hateful images or symbols in their profile image or profile header. Users may also not alter their username, display name, or profile bio in such a manner as to appear as if they are engaging in abusive behaviour, or which could be reasonably construed as causing harassment to other users or expressing hate. Examples of content that should not be posted, uploaded, streamed or shared include: • Content that includes names, symbols, logos, maps, flags, slogans or other objects related to hate and discrimination • Content promoting violence, hatred, segregation or discrimination • Content that supports misogyny, or hateful ideology against religion, caste, LGBTQ, or among others Under the section 'Violent and Graphic Content', Josh's Community Guidelines also prohibit users from using the platform to make any direct or indirect threats of physical harm against another person. This includes any threat relating to theft, vandalism, wrongful confinement, bodily, mental or financial harm. The platform does not permit its users to post content which involves references to mass murder, violent events, or specific means of violence, in particular where vulnerable communities or groups are the primary targets or victims, or which depicts or glorifies any such actions. Josh does not permit the glorification of violence in any manner including but not limited to celebrating events that have caused physical harm to or hurt

	the sentiments of any person or group. Users must refrain from posting content which depicts torture, injury, mental suffering, death, bodily harm, abduction, kidnapping or other forms of violent content. Josh allows content but with some limitations to help people raise awareness about issues. For example, users should not post, share, upload or stream: • Videos pertaining to people or dead bodies other than those in medical settings • Videos and photos that show the violence among each other • Content that shows the violent death of a person Under the section 'Illegal activities and regulated goods', Josh's Community Guidelines also prohibit the trade, sale, promotion, and use of certain regulated goods, as well as the depiction or promotion of criminal activities. Josh allows exception for content that provides value to the public, such as educational, scientific, artistic, and newsworthy content. Any content promoting or selling weapons, firearms, explosives, illegal goods or services is strictly prohibited. Lastly, Josh's Terms of Service prohibits: • intimidation or harassment of other people, or the promotion of violence or discrimination based on race, sex, religion, nationality, disability, sexual orientation or age • any material that contains a threat of any kind, including threats of physical violence • any material that would constitute, encourage or provide instructions for a criminal offence or dangerous activities • any material which is offensive, hateful or inflammatory • any material that is racist of discriminatory, including discrimination on the basis of someone's race, religion, age, gender, disability or sexuality • any material that is deliberately designed to provoke or antagonise people, or is intended to harass, harm, hurt, scare, distress, embarrass or upset people (Josh, 2023)
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://share.myjosh.in/terms-conditions?lang=en
3. Are there specific provisions applicable to livestreamed content in	Yes. Josh's ToS explain that Live Content may have different features and functionalities (e.g., sharing, commenting, interactivity features) from other types of user

the ToS or Community Guidelines/Standards?	content uploaded on the platform. And Josh's Live Content must respect the platform's ToS and Community Guidelines.
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?	Josh broadly states that it reserves the right to disable a user account, and remove or disable any content uploaded or shared by users, at any time, including if users have failed to comply with the ToS, or if activity on an account might cause damage to or impair the service, or infringe or violate any third party rights, or violate any applicable laws or regulations. In the event that multiple reports are made regarding violations of the ToS and Community Guidelines, Josh may terminate an account and block the user from registering again.
4.1 Notifications of removals or other enforcement decisions	Josh can remove any content in its sole discretion, with or without notice to users.
4.2 Appeal processes against removals or other enforcement decisions	Users can address their concerns to the grievance officer (grievance.officer@myjosh.in) regarding any matters related to platform usage, including appealing enforcement decisions.
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 or advertisement, either by email or via the complaint form. In order to be considered, complaints should include sufficient details such as the nature of the complaint, details about the content and the grievance made, the remedies sought, communication details, etc. Among the different reporting categories available, users can report content that 'hurts religious sentiment or incites violences', defined as 'any article/advertisement containing any image, text or any other content that hurts religious beliefs or sentiments of aggrieved party or directly incites or induces violent behaviour among people'. Content and advertisement containing hate speech, content inciting rebellion against the government or any religion or religious organisation can also be reported under this category. Once a content piece or a profile is reported for violation of the Community Guidelines, the content moderation team reviews each reported grievance, and depending on the severity of the violation takes action either in the form of content takedown or blocking of the account in accordance with Josh's Community Guidelines. Josh proactively moderates content. Content moderation on the platform is a three-step process. Josh's in-house machine learning product and processes are central to the

	review process, it helps detect and remove content that violates the Community Guidelines. Josh also uses robust third-party moderation tools that aid the moderation process. The next step is human review to evaluate violations and take further appropriate actions. Additionally, Josh indicates that it is strengthening its AI/ML capabilities to detect live content violations. Josh is not a member of the GIFCT.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Josh states that its policies describe permissible activities on the platform and these are reviewed and/or updated annually. Any user content that violates the policies is removed. In addition, depending on the severity of the violation, the user may also be blocked.
7. Does the service issue transparency reports (TRs) on TVEC?	Yes. Josh has published one transparency report covering January to September 2022 (Josh, 2022). Josh also reports separately on grievances (user reports, appeals, government requests and other requests) received monthly. The latest available information concerns July 2023 (Josh, 2023).

8. What information/fields of data are included in the TRs?	Josh's TR includes one TVEC-specific information:
	
	• Percentage of enforcement actions taken on content, broken down by category of violating content (as tagged by users), among which 'terrorism and extremism', 'violent and graphic content', and 'illegal activities and regulated goods'.
	
	Separately, Josh reports the following information on grievances on a monthly basis:
	
	Grievances and actions taken
	
	• Number of grievances pending at the beginning of the month, and number of number of grievances disposed out
	• Number of grievances received during the month, and number of grievances disposed out
	• Number of grievances pending at the end of the month
	
	Classification of grievances disposed
	
	• Number of grievances not related to Code of Ethics
	• Number of grievances related to Code of Ethics
	  ○ Agreed to by Josh and action taken
	  ○ Not agreed to by Josh

	○ Any other action taken Orders, directions and advisories received from Central Government and Self regulatory Bodies • Number of orders, directions and advisories received • Number of orders, directions and advisories complied with
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	The TR includes requests received by Josh from law enforcement agencies. Based on such requests, the actions taken include content/user account take down, and/or user data shared with the law enforcement agencies. The TR includes the user complaints received by Josh via the various reporting mechanisms (email and complaint form). The TR also provides definitions for each category of violating content: • Terrorism and Violent Extremism: Any content which threatens the security and national integrity of India or promotes any dangerous activities are a violation of the platform's policies. • Violent and Graphic Content: Any activity which directly or indirectly threatens physical harm against any person, related to theft, vandalism, wrongful confinement, bodily harm, violent events, etc. in the ways that can be construed as promoting or suggesting such content are a violation of the platform's policies. • Illegal Activities and Regulated Goods: The platform is committed to prohibiting and creating awareness about illegal activities and regulated goods like the display of weapons, arms, and ammunition and the promotion of criminal and dangerous activities.
10. Frequency/timing with which TRs are issued	Not specified. Josh states in its TR that it plans to 'periodically publish such reports'.
11. Has this service been used to post TVEC?	Yes. See section 8 above.
12. Main changes since last Report	Josh was not included in previous Reports.

42. 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. When necessary, we will cooperate with law enforcement authorities to handle such matters. 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. If Likee discovers content involving a risk of violence that may threaten public safety, it will ban the offending account. When necessary, it will cooperate with relevant government authorities. The following content is prohibited: • 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://likee.video/agreement and https://likee.video/community
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	Yes. Likee's ToS state that users may only use the live-streaming service if they are 18 years or older. Livestreamed content on the platform is also regulated by the UGC provision set forth in Section 4 of Likee's ToS.
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 may, but is not with obligation to, review, monitor, display, reject, refuse to post, store, maintain, accept or remove any content posted by the user. Upon detections, Likee may, in its sole discretion, delete, move, re-format, remove or refuse to post or otherwise make use of the content without any liability to the user or any third party in connection with its operation of Likee in an appropriate manner. Likee will also 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	Likee provides in-app notifications to users in real-time to inform them of content removal, account bans and feature bans.
4.2 Appeal processes against removals or other enforcement decisions	Users can appeal against removals or other enforcement decisions in some circumstances, such as: for account bans, device bans, permanent frozen of assets, livestream blockings, or livestream rejections due to the reason of age, users can appeal by clicking the “submit appeal” button or contacting Likee feedback services. Likee feedback services are always available for all communications.
5. Means of identifying TVEC (for example, monitoring algorithms, user-generated, human (staff) reviewers, hash-sharing/URL sharing database)	Likee uses a combination of AI and manual methods to review and monitor the publication of prohibited content, on a 24/7 basis. It has a content monitoring team and model algorithm team, who are responsible for the content environment of the platform. These teams review user feedback and reports. Likee also explains that it adopts strong and effective security strategies to proactively delete violating content from the platform. Likee reviews all videos reported by the AI system and its professional reviewing staff and takes reasonable measures to handle the reports. Content that violates the Community Guidelines is deleted, and certain content is subjected to specific further restrictions, e.g., age restriction control is imposed on content unsuitable for minors. Likee is not a member of the GIFCT.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Likee removes content that violates its Community Guidelines. Likee also penalises or bans user accounts involved in serious or repeated violations. If necessary, violations are reported to relevant legal authorities and Likee cooperates in investigations to ensure community safety. Likee also states that it may send warnings to users who publish violating content.
7. Does the service issue transparency reports (TRs) on TVEC?	No. However, Likee has published one ‘Violating Content Deletion Reports’, for the period going from April to June 2020. Although it does not contain TVEC-specific information, TVEC could possibly be included in the following reporting categories: ‘violent behaviour’, ‘harmful/dangerous behaviour’, ‘violence and gore’, ‘dangerous content’. No report has been published since. TRs are made depending on each official request for such disclosure. TRs data/information would then be extracted from

	automatic calculations exported from the content moderation mechanism.
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	Likee has published only one transparency report (April – June 2020).
11. Has this service been used to post TVEC?	Unknown.
12. Main changes since last Report	Likee has implemented a notification process and an appeal process.

43. 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 to 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 or describing hand signals representing gang affiliation. ○ Depicting or describing 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 upload, create, share, edit, search for or post violent content involving humans or animals, that is

	excessively bloody, vivid, gruesome, gory, or shocking. This includes content: ○ Depicting or describing disturbing footage of war, car crashes and other accidents. ○ Depicting or describing decapitations, suicide, terrorism, murder, or executions. ○ Depicting or describing wounds where the injury is the central focus. ○ Depicting or describing the torture, skinning, slaughter, mutilation, cruelty towards, or harm of animals or humans. ○ Depicting or describing 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 upload, create, share, edit, search for or 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. Picsart may permit certain content for historical, documentary, or educational purposes.
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. There is no livestream functionality on Picsart.
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 terminates accounts with severe or repeated violations (Picsart, n.d.). Determining whether there has been a violation of the 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	Users are notified in accordance with Picsart policies.

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 objectionable content on Picsart. All reports are reviewed by Picsart's team of moderators (Picsart, 2015). Picsart uses artificial intelligence in its content moderation efforts (Liao R. , 2019) In addition, Picsart relies on Space Owners and Admins to monitor content and behaviour on Picsart Spaces. Each Space provides its own rules and guidelines. Space Owners and Admins' responsibilities include but are not limited to monitoring posts, comments, promoted (pinned) content, Space description, and Space rules. In case of violation, they are required to promptly report the user profile or content. Space Admins may remove any content or community member that breaks the rules of a Space. However, Picsart reserves the right to make decisions it considers appropriate for the community (Picsart, 2023). Picsart is not a member of the GIFCT.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Picsart removes content that violates its Community Guidelines, Terms of Use, and other policies. Picsart may also restrict or terminate accounts that violate its policies. In certain circumstances, it may also report an account to the relevant authorities or law enforcement. In addition, Picsart reserve the right to revoke or limit a user's ability to create or administer a Space at any time in its sole discretion. If any Space Owner or Admin responds to communications from the Picsart team with hostility or if Picsart finds issues to be unresolvable via educational outreach, the following enforcement actions may be taken: • Temporary or permanent restriction of the user's access to Spaces or the Owner/Admin's own Space • Prohibiting a Space Owner/Admin from creating new Spaces • Prohibiting a Space Admin from becoming an Admin for other Spaces • Removal of content • Any other action Picsart deems appropriate based on the circumstances (Picsart, 2023)
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. Main changes since last Report	Picsart has a notification mechanism in place.

44. 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 (Tumblr, 2022) 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 dehumanises, individuals or groups based on race, ethnic or national origin, religion, gender, gender identity, age, veteran status, sexual orientation, disability or disease. Lastly, Tumblr prohibits violent content and threats, gore and mutilation; including encouraging or inciting violence, or glorifying acts of violence or the perpetrators.
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?	Yes, Tumblr Live was introduced in 2023 in several countries (Tumblr, 2023).[55] With this new feature, users can broadcast live streams, watch other streamers' broadcasts, and give or be given gifts. Tumblr Live is a service provided by The Meet Group (TMG), and governed by both Tumblr's and TMG's policies. Streams on Tumblr Live are moderated by TMG. Tumblr or TMG may suspend a user's access to Tumblr Live

	at any time if they believe that any applicable rules or terms have been violated (Tumblr, 2023). TMG's 'Content and Conduct Policy' prohibits violence and physical harm. Users may not make specific threats of violence or wish for the serious physical harm, death, or disease of an individual or group of people. This includes, but is not limited to, threatening or promoting terrorism (Tumblr Live, 2021).
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. From their account, in the 'Review flagged posts' section, users can see a timeline of their posts that have been flagged, appeal the classification, edit the post to remove flagged content, and check the status of their appeal.
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 trained experts from its 'Trust and Safety' team 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. For livestreamed content, users can submit a report to Tumblr's Live Support team, who will act quickly to review the content of the stream and take any necessary actions (Tumblr, 2023). Tumblr moderates content using a mix of machine-learning classification and human moderation from its team of trained experts. It notes that computers are better than humans at scaling process but they're not as good at making nuanced contextual decisions. Tumblr has observed that the 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.
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. Repeated violations of Tumblr's Community Guidelines may result in permanent block or account suspension.
7. Does the service issue transparency reports (TRs) on TVEC?	No. Oath, previous controller of Tumblr (Alexander J. , 2019), does release transparency reports. Up until the year 2018, they included Tumblr. However, the reports are very broad and do not break down the information per company controlled by Oath (for example, government requests for removal of content included both Yahoo and Tumblr). Also, there is no information specific to TVEC (Verizon Media, 2019). In 2019 Tumblr was sold to Automattic. Several Tumblr Transparency Reports have been published ever since, but none of them contain any information on TVEC (Tumblr, 2013-2022) Tumblr issued its first transparency report under the EU Regulation 2021/794 addressing the dissemination of terrorist content online (TCO), covering from 7 June 2022 to 30 June 2023. Since Tumblr has not received any removal order under the TCO for the reporting period, all fields indicate "0" (Tumblr, 2023).
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) (Fisher-Birch, 2018). Tumblr has ever since strived to improve its content moderation efforts, joining Tech Against Terrorism and the GIFCT.
12. Main changes since last Report	• Tumblr launched a livestream feature, with a dedicated Live Support team to review user reports of violating content. • Tumblr issued its first transparency report under the EU Regulation 2021/794 addressing the

	dissemination of terrorist content online (TCO), covering from 7 June 2022 to 30 June 2023.

45. Steam

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No definition is provided. Steam does not specifically prohibit TVEC, but generally prohibits users from the following:
	• Engage in unlawful activity, including encouraging real-world violence
	• Upload or post illegal or inappropriate content, such as real or disturbing depictions of violence
	• Harass other users or Steam personnel, including threatening
	
	In addition, Steam provides additional information in its 'Rules and Guidelines for Steam: Discussions, Reviews, and User-Generated Content' (Steam, s.d.). For each type of content, Steam gives a non-exhaustive list of examples of specific content and behaviour that are prohibited:
	• For Steam discussions and comment threads, the following is prohibited:
	  ○ Unlawful or prohibited activity such as encouraging or facilitating real-world violence
	  ○ Insults, harassment, threats or encouragement of harm
	• In user reviews, the following is prohibited:
	  ○ Threats or encouragement of harm
	  ○ Direct abuse or insults at other players, developers, or groups.
	• In user-uploaded content, the following is prohibited:
	  ○ Upload inappropriate content such as real or disturbing depictions of violence
	  ○ Content to harass, abuse, or otherwise disparage others
	
	Lastly, Steam allows players to create private or public community groups, each with their own discussion hub, announcement system, and group chatroom.

2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://store.steampowered.com/online_conduct?snr=, and https://help.steampowered.com/en/faqs/view/6862-8119-C23E-EA7B
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?	Steam moderators review reported content to determine when content should be removed from the community.
4.1 Notifications of removals or other enforcement decisions	Steam generally notifies content removals.
4.2 Appeal processes against removals or other enforcement decisions	If users think that their content has been mistakenly moderated, they can contact Steam Support to appeal the decision. Steam explains that support can review the content and help users better understand the reasoning or even correct a mistake.
5. Means of identifying TVEC (for example, monitoring algorithms, user-generated, human (staff) reviewers, hash-sharing/URL sharing database)	Steam encourages users to report any content that they believe to be in violations of Steam policies. In particular, they can report inappropriate content (user profiles/avatars, community groups, user-generated content), and abusive behaviour (Steam, 2023). Steam also relies on its community to moderate content in community groups. The creators and members of these groups are responsible for ensuring that they adhere to Steam guidelines. Steam is not a member of the GIFCT.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	In case of violation of Steam's policies, Steam may take the following actions: • Content removal: Content in violation of the guidelines will be removed from Steam. Generally, users will receive a notification if this occurs, with further information and an option to contact support. • Cooldown for specific features: The ability to use community features relevant to the banned content may be temporarily removed from the account. For example, if an avatar is removed, the account will lose the ability to upload a new avatar until the cooldown has expired. • Loss of all community features: Accounts that repeatedly violate the guidelines risk temporarily or permanently losing the ability to use all features within the Steam Community.

	Admins and moderators have a range of tools at their disposal to moderate content within Steam Hub Communities (Steam, s.d.).
	
	Discussion moderation tools:
	
	• Customising subforums:
	  ○ Add or remove subforums
	  ○ Control who can view each subforum and the content in it
	  ○ Control who is able to post in existing threads and creating new ones
	• Editing posts
	• Deleting threads or posts
	• Merging, pinning, locking, or marking as answered threads
	
	Warning, bans and messaging:
	
	• Warnings: Moderators can use pre-written messages are translated into whatever language the player has Steam set to; or write a custom warning reason. Warning is the first response to players breaking the Steam Community Rules and Guidelines.
	• Bans: Bans are a last resort for violations of the Steam Community Rules and Guidelines. They should be temporary in order to serve as a reminder of the discussion rules. If a player has received previous temporary bans, it may be appropriate to permanently ban them from the Hub. In this case, moderators should leave a clear, concise ban reason. This ban reason is visible to the player as well as other moderators. Banned accounts will not be able to interact with the Community Hub, including creating new threads and posting in existing ones. Banned accounts are also not allowed to upload screenshots, artwork, guides, workshop items or other user-generated content.
	
	Moderating screenshots, videos, and workshop items:
	
	• Hiding content as inappropriate: This option is used to hide the content for players who have asked not to be exposed to extreme violence or sexual content. Hiding as inappropriate will blur the preview image for the content, preventing players from seeing it without acknowledgment of an age-gate.
	• Banning items: Banning content should be reserved for items that violate the Steam Subscriber Agreement.

	Players will receive an email that their item has been banned and that it is only visible to them. **Moderating reviews:** • Flagging a review • Responding to a review
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. Steam has become popular for White supremacists. The extreme right uses Steam as a hub for individual extremists to connect and socialise (Lakhani, Video Gaming and (Violent) Extremism: An exploration of the current landscape, trends, and threats, 2021).
12. Main changes since last Report	Steam was not included in previous Reports.

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. Since the last report, Google updated its Abuse Program Policies (Google, n.d.), modifying the different categories of violating content. TVEC is now included in the 'Violent Organisations and Movements' category, whereas before it was in a separate category titled 'TerroristaActivities'. In substance, the same content and behaviours are still prohibited but the scope of this particular category extends beyond just TVEC. Other relevant categories of violating content include Violence and Gore, Hate Speech, and Dangerous and Illegal Activities. **Violent Organisations and Movements:** Known violent non-state organisations and movements are not permitted to use Google Drive for any purpose. Users must not distribute content that facilitates or promotes the activities of these

	groups, such as recruiting, coordinating online or offline activities, sharing manuals or other materials that could facilitate harm, promoting ideologies of violent non-state organisations, promoting terrorist acts, inciting violence, or celebrating attacks by violent non-state organisations. If users post content related to violent non-state organisations for an educational, documentary, scientific, or artistic purpose, they must provide enough information so viewers understand the context. **Violence and Gore:** Users must not store or distribute violent or gory content involving real-life people or animals that is primarily intended to be shocking, sensational, or gratuitous. This includes ultra-graphic violence, such as dismemberment or close-up footage of mutilated corpses. Also, users must not encourage others to commit specific acts of violence. Graphic material, such as content containing significant amounts of blood, may be allowed in an educational, documentary, scientific, or artistic context, if sufficient context is provided. In some cases, content may be so violent or shocking that no amount of context will allow that content to remain on Google Drive. **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. **Dangerous and Illegal Activities:** Users must not use Google Drive to engage in illegal activities or to promote activities, goods, services, or information that cause serious and immediate harm to people or animals. While Google Drive permits general information for educational, documentary, scientific, or artistic purposes about this content, content that directly facilitates harm or encourages illegal activity is prohibited. Google will take appropriate action if it is notified of unlawful activities, which may include reporting users to the relevant authorities, removing access to some of its products, or disabling an account.
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. There is no livestream functionality on Google Drive.
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 (Google, n.d.).
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 the service 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. A file review typically takes around 5 days. The exact time frame depends on the nature of the review request.
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. The available categories for reporting a violation include, among others, terrorism, violence, hate speech, and illegal activities. 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). To enforce its policies, Google Drive relies on a combination of automated and human efforts to identify potentially violative content. Users can report potentially violative content via in-product tools. In addition, Drive's automated systems are trained to quickly identify violative content. This includes flagging potentially problematic content

	for human reviewers, whose judgement is needed for the many decisions that require a more nuanced determination. Google Drive is owned by Google, which joined the GIFCT in 2022.
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 • Report illegal materials to appropriate law enforcement authorities (Google, n.d.)
7. Does the service issue transparency reports (TRs) on TVEC?	No. Google issues semi-annual TRs (Google, n.d.) encompassing Google's products and services, including Google Drive. These reports contain a section on government requests to remove content based on violations of local laws or Google's ToS or policies, but there is no TVEC-specific information. The last available report covers July to December 2022.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Has this service been used to post TVEC?	Yes. ISIS content has been found on Google Drive (Katz, To Curb Terrorist Propaganda Online, Look to YouTube. No, Really., 2018).
12. Main changes since last Report	• Google modified its Abuse Program Policies, that apply to Google Drive. The category of violating content titled 'Terrorist activities' does not exist anymore and TVEC is now included in the broader 'Violent Organisations and Movements' category. In substance, the same content and behaviours are still prohibited but the scope of the new category extends beyond just TVEC. However, the list of available categories of

	violating content for user reporting still mentions a 'terrorist content' category. • Google, Google Drive's parent company, officially joined the GIFCT.

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, share or store content that contains or promotes extreme acts of violence or terrorist activity, including terrorist or violent extremist propaganda. Using Dropbox to advocate bigotry, hatred or the incitement of violence against any person or group of people based on their race, religion, ethnicity, national origin, sex, gender identity, sexual orientation, disability, impairment or any other characteristic(s) associated with systemic discrimination or marginalisation 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/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, it will suspend or terminate the user's access.
4.1 Notifications of removals or other enforcement decisions	In cases of breaches of its ToS, 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).
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.). 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 (now X), 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). 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. Dropbox reserves the right to take appropriate action in response to violations of its Acceptable Use Policy, which could include removing or disabling access to content, suspending a user's access to the Services or terminating an account.
7. Does the service issue transparency reports (TRs) on TVEC?	No. Dropbox issues semi-annual TRs (Dropbox, n.d.) that contain a section on government requests to remove content based on violations of local laws or Dropbox's ToS or policies, but there is no TVEC-specific information. The latest available report covers from July to December 2022.
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).
12. Main changes since last Report	No main changes since last Report.

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 the user, the service, or 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) that depicts graphic violence, encourages violent action, endorses a terrorist organisation or its acts, or encourages people to join such groups. The UN Sanctions List includes a list of groups that the UN Security Council considers to be terrorist organisations (Microsoft, 2016).
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 Service Agreement. When investigating alleged violations of its Services Agreement, 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).
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 users have the opportunity to appeal account actions by visiting its appeals webpage (https://www.microsoft.com/en-us/DigitalSafety/moderation-and-enforcement/appeals) or using its appeals web from (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 has a dedicated reporting mechanism through which users can report 'terrorist content posted to a Microsoft consumer service'. Microsoft encourages users to use this form to report content posted by or in support of a terrorist organisation that depicts graphic violence, encourages violent action, endorses a terrorist organisation or its acts, or encourages people to join such groups (Microsoft, 2023). Microsoft deploys a variety of scanning technology, artificial intelligence, external partnerships, and human moderation operations solutions to detect and investigate TVEC. In particular, Microsoft explains that to detect potential child sexual exploitative content and/or terrorist and violent extremist content, it uses scanning technologies (e.g., PhotoDNA or MD5) and other AI-based technologies, such as text-based classifiers, image classifiers, and the grooming detection technique. 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). Microsoft is a founding member of the GIFCT and chairs the GIFCT Operating Board. Via the GIFCT, Microsoft participates in a range of activity, including engagement in its multi-stakeholder working groups and the GIFCT's Incident Response processes. In the event the GIFCT and its Operating Board activate a Content Incident or Content Incident Protocol, Microsoft ingests

	related hashes from the GIFCT's hash-sharing database. This allows Microsoft to quickly become aware of, assess, and address potential content circulating on its consumer services resulting from an offline terrorist or violent extremist event.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	See information in Section 4 above. 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)
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, 2020-2022). The latest transparency report covers the period from July to December 2022. 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. Microsoft also publishes an annual report under Regulation (EU) 2021/784 on addressing the dissemination of terrorist content online. The latest report is available on its CSR Trust Hub (https://www.microsoft.com/en-us/corporate-responsibility/reports-hub#coreui-feature-jy1t3q4) and Jurisdictional transparency reports page (https://www.microsoft.com/en-us/DigitalSafety/transparency-reports/jurisdictional-reports).
8. What information/fields of data are included in the TRs?	The latest transparency report (July - December 2022), includes the following information: • Number of TVEC content actioned • Percentage of TVEC content detected proactively • Number of accounts actioned due to TVEC • Percentage 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 actioned' refers to when Microsoft suspends or blocks access to an account, or restricts access to content within the account. 'Proactive detection' refers to Microsoft-initiated flagging of content on its services, whether through automated or manual review. 'Accounts reinstated" refers to actioned accounts that were fully restored including content and account access, upon appeal.
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).
12. Main changes since last Report	Refreshed language on Microsoft's practices, processes and systems on addressing TVEC. Microsoft's also launched its new Digital Safety site (https://www.microsoft.com/en-us/DigitalSafety) with further details on its efforts to address TVEC.

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, under its 'Terrorist activity' policy, 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, 2023), with which WordPress.com is prohibited by law from doing business. WordPress.com does not allow individuals, groups, or entities on that list to use WordPress.com (Word Press, n.d.). Genuine calls to violence are also prohibited. This 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. Lastly, Wordpress.com's User Guidelines prohibit direct and realistic threats of violence. Users cannot post a genuine call for violence – or death – against an individual person, or

	groups of persons. This does not mean that all hyperbole or offensive language will be 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 Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	WordPress.com has worked in conjunction with experts on online extremism, as well as law enforcement, to develop policies to address extremist (not specifically violent extremist) and terrorist propaganda. WordPress.com suspends websites that call for violence or that are connected to officially banned terrorist groups (per the US Treasury's OFAC list), regardless of content. WordPress.com also 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 (Clicky, 2017).
4.1 Notifications of removals or other enforcement decisions	WordPress.com states that, depending on the scenario, it will email or add a warning notification in the dashboard of a user violating its policies. The notification will contain a link that the user can use to contact WordPress.com regarding the issue. However, those 'scenarios' are not specified (WordPress.com, 2023).
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 human 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 can report content or sites in violation of WordPress.com's policies. 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). The EU Regulation 2021/784 addressing the dissemination of terrorist content online (TCO) entered into force on 7 June 2022. Automattic notes that, while it is legally required to remove all content that is subject of an EU Regulation content removal order, it still manually reviews each order to ensure accuracy and validity (Automattic, 2023). 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, 2023). The last transparency report includes data from 1 July to 31 December 2022. Worpress.com issued its first transparency report under the EU Regulation 2021/794 addressing the dissemination of terrorist content online (TCO), covering from 7 June 2022 to 30 June 2023. Since Wordpress.com has not received any removal order under the TCO for the reporting period, all fields indicate "0" (Automattic, 2023).
8. What information/fields of data are included in the TRs?	The transparency report features the following metrics in relation to IRU reports: • Number of IRU notices received • Percentage of notices where sites/content were removed as a result • Number of sites/content specified The figures are broken down by month and by reporting entity (e.g. Europol) or country. It is not specified which categories of violating content are included in the reported metrics. However, in the Summary section, Automattic explains that terrorist and extremist propaganda is a particular focus, and that IRUs have expertise in identifying sites used by known terrorists to spread propaganda or to organise acts of violence. In addition, the transparency report features the following metrics in relation to the EU terrorist content removal orders: **Removal orders** • Number of EU Regulation content removal orders received • Number of items removed or blocked as the result of removal orders • Number of removal orders where content was not removed or blocked **Appeals** • Number of user appeals for content removed as the result of a removal order

	• Number of user appeals for content removed as a result of our standard moderation practices • Number of EU Regulation removal orders appealed by Automattic **Reinstatement** • Number of cases where Automattic reinstated content previously targeted by an EU Regulation order based on successful user appeal • Number of cases where Automattic reinstated content previously removed by standard moderation practices based on successful user appeal • Number of cases where administrative or judicial review proceedings resulted in reinstatement of content As of 31 December 2022, Wordpress.com has not received any EU Regulation content removal orders, therefore, all the reported numbers are 0.
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 7 above. For instance, in 2018, the Referral Action Days – a joint campaign with the participation of Europol's European Union Internet Referral Unit, and the national referral units of Belgium, France the Netherlands, Slovenia and the United Kingdom – targeted terrorist content located on Wordpress.com and VideoPress, a video-hosting service for WordPress sites. As a result, more than 900 items of banded terrorist propaganda were swiftly flagged to the platform moderators for further review and eventual removal (EUROPOL, 2018).
12. Main changes since last Report	• Worpress.com issued its first transparency report under the EU Regulation 2021/794 addressing the dissemination of terrorist content online (TCO), covering from 7 June 2022 to 30 June 2023.

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. In addition, the Wikimedia Foundation has a Universal Code of Conduct (Wikimedia Foundation, 2023) that applies to all of its projects, including Wikipedia. The section titled 'Unacceptable behaviour' of this Code prohibits: - Threats: Explicitly or implicitly suggesting the possibility of physical violence, unfair embarrassment, unfair and unjustified reputational harm, or intimidation by suggesting gratuitous legal action to win an argument or force someone to behave the way you want. - Encouraging harm to others: This includes encouraging someone else to commit self-harm or suicide as well as encouraging someone to conduct violent attacks on a third party. - Hate speech in any form, or discriminatory language aimed at vilifying, humiliating, inciting hatred against individuals or groups on the basis of who they are or their personal beliefs. - The use of symbols, images, categories, tags or other kinds of content that are intimidating or harmful to others outside of the context of encyclopaedic, informational use. This includes imposing schemes on content intended to marginalise or ostracize. - Insults: This includes name-calling, using slurs or stereotypes, and any attacks based on personal characteristics. Insults may refer to perceived characteristics like intelligence, appearance, ethnicity, race, religion (or lack thereof), culture, caste, sexual orientation, gender, sex, disability, age, nationality, political affiliation, or other characteristics. In some cases, repeated mockery, sarcasm, or aggression constitute insults collectively, even if individual statements would not.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://foundation.wikimedia.org/wiki/Terms_of_Use/en, https://en.wikipedia.org/wiki/Wikipedia:Policies_and_guidelines#Enforcement and https://foundation.wikimedia.org/wiki/Policy:Universal_Code_of_Conduct
3. Are there specific provisions applicable to livestreamed content in the	Not applicable.

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?	The Wikipedia community has the primary role in creating and enforcing its policies. The community is composed of: • *Editors*: volunteers who write and edit the pages of Wikipedia • *Stewards*: volunteer editors tasked with the technical implementation of community consensus, with Checkuser (Wikipedia, 2019) and oversight (Wikipedia, 2020) powers. • *Bureaucrats*: volunteer editors with the technical ability (user rights) to promote other users to administrator or bureaucrat status, remove the admin status of other users, and grant and revoke an account's bot status. • *Administrators*: editors who have been trusted with access to restricted technical features ("tools"). For example, administrators can protect and delete pages, and block other editors (Wikipedia, 2020). Wikipedia's administrators can perform certain special actions. These include the ability to block and unblock user accounts, IP addresses, and IP ranges from editing, edit fully protected pages, protect and unprotect pages from editing, delete and undelete pages, rename pages without restriction, and use certain other tools (Wikipedia, 2023). Wikipedia's core content policies are: 1. Neutral point of view: All Wikipedia articles and other encyclopaedic content must be written from a neutral point of view, representing significant views fairly, proportionately and without bias. 2. Verifiability: It means that people reading and editing the encyclopaedia can check that information comes from a reliable source. 3. No original research: Wikipedia does not publish original thought. All material in Wikipedia must be attributable to a reliable, published source (Wikipedia, 2019). Content is deleted by the administrators if it is judged to violate Wikipedia's content or other policies, or the laws of the United States (Wikipedia, 2020). The deletion process encompasses the processes involved in implementing and recording the community's decisions to delete

	pages and media (Wikipedia, 2020). Normally, a deletion discussion must be held to form a consensus to delete a page. In general, administrators are responsible for closing these discussions, though non-administrators in good standing may close them under specific conditions. However, editors may propose the deletion of a page if they believe that it would be an uncontroversial candidate for deletion. In some circumstances, a page may be speedily deleted if it meets strict criteria set by consensus, which include pages that disparage, threaten, intimidate or harass their subject or some other entity, and serve no other purpose (Wikipedia, 2020). The Wikimedia Foundation states that it rarely intervenes in community decisions about policy and its enforcement. However, when the community requires intervention, or to address an especially problematic user because of significant disturbance or dangerous behaviour, the Wikimedia Foundation may investigate the user's use of the service (a) to determine whether a violation of any policies or laws has occurred, or (b) to comply with any applicable law, legal process, or appropriate governmental request. After the investigation, sanctions may be applied (see Section 6 below). Finally, Wikipedia's Arbitration Committee is the panel of editors and administrators, elected by the community at large, responsible for conducting arbitration process. It has the authority to impose binding solutions to disputes between editors, primarily for serious conduct disputes the community has been unable to resolve, as a last resort process. It must be noted that it does not make decisions relating to content. The Arbitration Committee refers requests regarding 'threats of harm to self or to others' to the Wikimedia Foundation emergency address (Wikipedia, 2023).
4.1 Notifications of removals or other enforcement decisions	Administrators should notify users when blocking them by leaving a message on their user talk page and supply a clear and specific reason why a user was blocked. When they are unblocked, users will also be notified of unblock conditions on their talk page (Wikipedia, 2023).
4.2 Appeal processes against removals or other enforcement decisions	Bans may be appealed to the community, the Arbitration Committee, or the Wikimedia Foundation, depending on the nature of the ban. Only individuals directly involved in a case may request review, either as requester or as an individual under investigation. When submitting an appeal, users should include an explanation as to why they believe the case should have been handled differently. The Interim Trust & Safety Case Review Committee, composed of ten experienced volunteers from the Wikimedia community, reviews appeals for eligible Trust & Safety office actions. The appealed cases are handled in the following way:

	• Appeal eligibility determined: The review requester will be notified whether or not a case is eligible for review. • Investigation: If the appealed case is eligible for review, the committee chair will assign five committee members to review the case. The committee will then review the case over a period of 14 business days. • Result: The review requester will be informed of the committee's decision. The committee does not have the authority to issue sanctions itself, but may overturn eligible Foundation office action decisions where they deem appropriate or send a case back to Trust & Safety for further investigation (Wikipedia, 2023).
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). In addition, the Wikimedia Foundation has a Trust and Safety team that handles reports of major safety issues on Wikimedia projects including suicide threats, threats of violence, and child pornography (Wikipedia, 2023). In addition, the Moderator Tools team is a Wikimedia Foundation product team working on content moderation tools. The team's focus is on content moderation processes, including page protection, deletion, reporting, and recent changes patrolling, rather than user reporting and moderation. In 2023, it started working on an Automoderator tool for Wikimedia projects that would allow moderators to configure automated prevention or reversion of bad edits based on scoring from a machine learning model (Wikipedia, 2023). In the particular case of 'threats of harm', the Wikimedia Foundation has an emergency address for users to report threats of physical harm only. Wikipedia states that generally, specific threats should be removed by administrators (Wikipedia, 2023). Wikipedia is not a member of the GIFCT.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	In case of violation of Wikipedia's guiding documents, the Wikipedia community or the Arbitration Committee may take two types of sanctions: • General sanctions (which apply to all editors in a particular topic area and empower administrators to sanction editors

	who are not complying with general behavioural or editorial guidelines and policies) (Wikipedia, 2023): ○ Contentious topics restrictions: Rapid measures for administrators to intervene in topic areas that have proved problematic, including but not limited to blocks of up to one year, article or topic bans, and revert restrictions. For specific pages, administrators may also impose preventative restrictions such as page protection and revert restrictions. ○ Page restrictions: Pages may be subject to restrictions that limit the types of edits that may be made. ○ Extended confirmed restriction: Limits editing in a specific topic area to only those accounts that have been extended confirmed. • Personal sanctions / editing restrictions (which only apply to individual editors) (Wikipedia, 2023): ○ Account restriction: The user is limited to editing with a certain number of accounts (usually one). ○ Civility restriction: The user may be sanctioned (including blocks) if they make any edits which are judged by an administrator to be uncivil, personal attacks, or assumptions of bad faith. ○ Probation (supervised editing): The user on probation may be banned from pages that they edit in a certain way (usually disruptively) by an uninvolved administrator. Probation is usually used as an alternative to an outright topic ban in cases where the editor shows some promise of learning better behaviour. ○ Move ban: The user is prohibited from directly moving (renaming) pages with the page move feature (sometimes relating only to specific topics or namespaces). ○ Revert restriction: The user is limited to a certain number of reverts (usually one) per page/topic per period of time (usually 24 hours or one week); exceptions, such as obvious vandalism, may apply. The user is additionally required to discuss any content reversions on the page's talk page. ○ Topic ban: The user is prohibited from editing either (1) any page (or section of a page) relating to a particular topic, (2) particular pages that are specified in the ban, and/or (3) making any edits in relation to a particular topic. Such a ban may include corresponding talk pages. ○ Article ban or page ban: The user is prohibited from editing a specific page or set of pages specified in

	the ban. Such a ban may include corresponding talk pages. ○ Interaction ban: The user is prohibited from interacting with one or more users. Generally, blocks for violating a civility restriction, a revert limitation, or a topic ban start at 24 hours per violation. Editors who violate Arbitration-imposed restrictions may be blocked or otherwise sanctioned. Finally, the Wikimedia Foundation may refuse, disable, or restrict access to the contribution of any user who violates its ToS, delete pages created by, or 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 semi-annual TRs (Wikimedia Foundation, 2022) covering requests for user data and requests for content alteration and takedown, but there is no section specifically addressing TVEC. The latest available report covers from July to December 2022.
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 voluntary 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.
12. Main changes since last Report	• In 2023, Wikimedia's Moderator Tools team started working on an Automoderator tool for moderating content automatically, based on machine learning.

Annex C. The Global Top 50 TVEC-intensive Services[56]

Rank	Name of service (parent company)	Type of service	Issues TVEC transparency reports	Provided feedback / comments on its profile
1	Telegram (Telegram Messenger LLP)	Messaging app	N	N
2	X (X Corp.)	Short messages-focused social networking platform	N	Y
3	Facebook (Meta, Inc.)	Social networking platform	Y	Y
4	Instagram (Meta, Inc.)	Social networking platform	Y	Y
5	TikTok (ByteDance Technology Co.)	Short video app	Y	Y
6	Rocket.Chat (Rocket.Chat Technologies Corp.)	Messaging app	N	N
7	Rumble (Rumble Inc.)	Video streaming platform	N	N
8	Discord (Discord, Inc.)	Chat platform	Y	N
9	LiveGore.com (Unknown)	Video streaming and publishing website	N	N*
10	ChirpWire (Unknown)	Social networking platform	N	N
11	Element (New Vector Ltd)	Messaging app	N	Y
12	Gab (Gab AI, Inc.)	Social networking platform	N	N
13	Mastodon (Mastodon gGmbH)	Social networking and microblogging platform	N	N
14	TamTam.Chat (VK LLC)	Messaging app	N	N
15	Matrix (New Vector Ltd)	Decentralised communication service	N	Y
16	Abolitionmedia.noblogs.org (Unknown)	Anonymous blogging platform	N	N*
17	Americanfuturistpublishing.com	Anonymous blogging platform	N	N*
18	Malhm.xyz (Unknown)	Content-sharing website	N	N*
19	Alazaimll.websites.co.in (Unknown)	Content-sharing website	N	N*
20	Umarmediattp.org (Unknown)	Content-sharing website	N	N*
21	Shadanews.com (Unknown)	News aggregation website	N	N*
22	4chan (4chan community support LLC)	Imageboard platform	N	N
23	Nuceciwan127.xyz (Unknown)	News aggregation website	N	N*

24	Amjaad.video (Unknown)	Video-sharing website	N	N*
25	Dalelansar.info (Unknown)	Content-sharing and news aggregation website	N	N*
26	Ansarollah.com (Unknown)	Content-sharing and news aggregation website	N	N*
27	Alqassam.ps (Unknown)	Content-sharing and news aggregation website	N	N*
28	Threads (Meta Platforms, Inc.)	Text-based social media app	N	N
29	GoyimTV.com (Unknown)	Video streaming platform	N	N
30	Moqawama.ir (Unknown)	Content-sharing website	N	N*
31	3pdirectory.com (Unknown)	Content-sharing website	N	N*
32	Odysee (Odysee, Inc.)	Video streaming platform	N	Y
33	YouTube (Alphabet, Inc.)	Video streaming platform	Y	Y
34	Archive.org (The Internet Archive, a 501(c)(3) non-profit Internet library)	Internet library	N	N
35	Justpaste.it (Wise Web Mariusz Żurawek)	Anonymous pastebin site	Y	Y
36	Google Drive (Alphabet, Inc.)	Cloud-based file sharing	N	Y
37	SoundCloud (SoundCloud Global Limited & Co. KG)	Audio streaming platform	N	Y
38	Dropbox (Dropbox, Inc.)	Cloud-based file sharing	N	N
39	MediaFire (MediaFire, LLC)	Cloud-based file sharing	N	N
40	Telegraph (Telegram Messenger LLP)	Anonymous blogging platform	N	N
41	Itarchives.org (Unknown)	File sharing	N	N
42	DoxBin.org (Unknown)	File sharing and publishing website	N	N
43	File.io (Mr Cowboy LLC)	Anonymous file sharing	N	N
44	Pixeldrain (Fornaxian Technologies)	File sharing	N	N
45	Gofile.io (Wojtek SAS)	File sharing	N	N
46	Signal (Signal Messenger LLC)	Messaging app	N	N
47	Wire (Wire Swiss GmbH)	Messaging app	N	N
48	Slack (Slack Technologies, LLC)	Cloud-based team communication platform	N	N
49	WhatsApp (Meta, Inc.)	Messaging app	N	Y
50	Threema (Threema GmbH)	Messaging app	N	Y

* In cases where no functional email address or contact form was available for a Service listed in Annex C, or where the Service appeared to be dedicated or sympathetic to terrorism and/or violent extremism, the OECD did not send a feedback request for the associated profiles. These Services are marked “N*” in the right-most column.

Annex D. Profiles of the Top 50 TVEC-intensive

1. Telegram

See profile 12 in Annex B.

2. X

See profile 21 in Annex B.

3. Facebook

See profile 1 in Annex B.

4. Instagram

See profile 3 in Annex B.

5. TikTok

See profile 7 in Annex B.

6. Rocket.Chat

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	No definition is provided. Rocket.Chat's Acceptable Use Policy (Rocket.Chat, 2023) states that users may not use the Services to provide material support or resources (or to conceal or disguise the nature, location, source, or ownership of material support or resources) to any organisation(s) designated by the United States government as a foreign terrorist organisation pursuant to section 219 of the Immigration and Nationality Act or other laws and regulations concerning national security, defence or terrorism. In addition, Rocket.Chat prohibits posting or sharing "harmful content", which includes content that: • Encourages any illegal activity including but not limited to, terrorism, inciting racial hatred, or the submission of

	which in itself constitutes committing a criminal offence; • Is obscene, pornographic, violent or otherwise may offend human dignity, or contains nudity; • Contains language or imagery which could reasonably be deemed offensive, or is likely to harass, upset, embarrass, or annoy any other person (including, but not limited to, any sort of language or imagery that could be deemed discriminatory against any race, religion, gender identity, sex, sexual orientation, colour, ethnicity, national origin, or ability status); • Is abusive, insulting, or threatening, discriminatory, or promoting of or encouraging racism, sexism, hatred or bigotry. Finally, Rocket.Chat's Code of Conduct (Rocket.Chat, 2019) states that it does not allow the following behaviour: • Posting of links (URLs) to offensive material, sites hosting malware, initiating downloads or promoting illegal activities • Offensive, rude, disruptive and unwanted • Posting of pornographic / gore / nude images • Harassment, name-calling, racist or sexist remarks, profanity towards others and other forms of bullying
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://docs.rocket.chat/applicable-terms/supplemental-terms/acceptable-use-policy and https://docs.rocket.chat/applicable-terms/supplemental-terms/terms-of-use
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?	Rocket.Chat does not assume any responsibility for the content generated by Users, including social media pages, video-sharing platforms, bulletin boards, discussions, and channel rooms. Users are solely responsible for any content created and Rocket.Chat does not endorse or guarantee the accuracy or legality of such content in any way. Rocket.Chat strongly recommends users seeking professional or specialist advice before making any decisions or taking any actions based on the content on its platform. In addition, Rocket.Chat may contain links to other websites or platforms and resources provided by third parties. Rocket.Chat

	explains that these links are provided for informational purposes only and should not be interpreted as endorsement or approval of the linked sites or platforms or the information the User may obtain from them. Rocket.Chat states that it does not have control over the content of those sites, platforms, or resources, and therefore is not responsible for any content, products, services, or actions of those third-party sites or platforms. It is the User's responsibility to exercise caution and use their own judgement when accessing and using any external links provided through the services (Rocket.Chat, 2023). If Rocket.Chat, in its sole discretion, determines that a policy violation is deliberate, repeated, or presents a credible risk of harm to other Users, its Customers, the Services, or any third parties; it may suspend or terminate access to the Services without prior notice and without any liability to concerned parties. Rocket.Chat states that it complies with valid local or international law enforcement requests to remove content or produce user data.
4.1 Notifications of removals or other enforcement decisions	Rocket.Chat informs the users affected by a law enforcement request to remove content or produce user data.
4.2 Appeal processes against removals or other enforcement decisions	If a user believes that an administrator has misused their authority, they can contact support@rocket.chat, providing the following information: • Administrator's open.rocket.chat nickname; • What is believed to demonstrate an abuse of power; • Associated dates and times (required to review log files etc.) (Rocket.Chat, 2019)
5. Means of identifying TVEC (for example, monitoring algorithms, user-generated, human (staff) reviewers, hash-sharing/URL sharing database)	Rocket.Chat states that it wants to be a platform that allows for free and unrestricted communication. It does not plan or want to build any kind of backdoor, censorship tool, or hidden remote control mechanism into Rocket.Chat. Users can report harmful content through the dedicated feature and Rocket.Chat will look into removing it. If users encounter another Rocket.Chat instance that is not hosted by Rocket.Chat and which they think contains illegal or otherwise harmful content, Rocket.Chat recommends reaching out to the administrator of that instance to moderate the related content. As an ultimate resort, Rocket.Chat advises users to contact the law enforcement body in charge of investigating the potential

	offence in question, to know the legal remedies available and the potential next steps to take (Rocket.Chat, 2020).
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Administrators are responsible for configuration and content moderation decisions within their instance. They can use the following features to moderate content: • Making use of the “moderator”-permission in channels to appoint individuals to purge or modify inappropriate messages • Notification feature for the use of specified words of phrases • Blacklisting certain words or phrases • Notifying users of applicable policies via e.g. pinning messages or adding an announcement to the room • Requiring confirmation of user registration by an administrator, to prevent unvetted users from posting messages • Enabling or disabling end-to-end encryption: with end-to-end encryption enabled, only an encrypted string of the message is stored on the server. This however prevents content auditing via administrators and moves responsibility for content moderation to users. • Turning on GoogleVision integration for image uploads, which has options to block images containing graphic or adult content All of these features are optional so administrators have the total flexibility in what to apply in their specific case. If a user is asked to modify their behaviour or given some other form of instruction but fail to comply, they are likely to be removed from the channel, the discussion or the whole server for a discretionary or unlimited period of time depending on the severity of the matter, depending on the administrator's decision. Depending on the severity of the misbehaviour, users can get banned without warning or be issued a warning. Should administrators deem a subject matter to be inappropriate for the channel and / or the current audience, they may at their discretion make contact via Private Message, making clear any requirement to cease discussing the topic, or - as per their permission level - ban the user. A ban may be temporary (in form of a ban period) or permanent. In addition, the server administrator has the right to remove your user account. When being corrected by an administrator, users should not argue with them in the main chat. They should request a PM session to help avoid conflict and keep the involvement of others to a minimum.

	Attacking administrators in a channel will likely result in a ban. Any genuine visitor should respect their decisions - they have nothing to gain from issuing bans / mutes and are doing so to the best of their ability, on behalf of the community. Administrators may also choose to request users to leave if they are deemed to be a risk to the safety of the general users / guests within the channel. Not adhering to the request will result in manual removal. Rocket.Chat states that when it receives requests from law enforcement to remove content, in most cases, it cannot remove the majority of content, because it is outside of its control on servers it does not have (and does not want) access to. If the content in question is on its Open Server, Rocket.Chat removes it if it is a breach of its code of conduct or it is compelled by a law enforcement request. For servers hosted by Rocket.Chat and under control of its customers, it removes content after notifying and in collaboration with the customer or directly as a violation of its ToS. Requests for the content of communications (e.g., messages, files) require a valid search warrant or equivalent from an agency with proper jurisdiction over Rocket.Chat (Rocket.Chat, 2020).
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. Main changes since last Report	Rocket.Chat was not included in previous Reports.

7. Rumble

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 Content Policies (Rumble, 2023).
	
	Such prohibited content includes:
	
	• Content or material that is grossly offensive to the online community, including but not limited to, racism, antisemitism and hatred
	• Content or material that:
	  ○ Promotes, supports, or incites violence or unlawful acts,
	  ○ 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/or
	  ○ Promotes or supports entities and/or persons designated by either the Canadian or United States government as terrorists or terrorist organisations.
	
	In 2022, Rumble issued a press release outlining a proposed new content policy and removal and appeal process (collectively the "Rumble Rules"), available at https://corp.rumble.com/blog/rumble-proposes-an-open-source-content-moderation-policy-process-to-improve-transparency-put-creators-first/
	
	Through phase 1, Rumble welcomed feedback from Rumble creators and users on the proposed Rumble Rules. However, no additional information about next phases or potential implementation has been published since the last Report.

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	There is a livestream functionality, but no specific provisions governing it.

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?	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). 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.
11. Main changes since last Report	• No main changes since the last Report. • In 2022, Rumble published a proposal for updating its "Rumble Rules", including a new content policy as well as new removal and appeal processes. Through phase 1, Rumble welcomed feedback from Rumble creators and users on the proposed Rumble Rules. However, no additional information about next phases or potential implementation has been published since the last Report.

8. Discord

See profile 21 in Annex B.

9. LiveGore.com

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	LiveGore Terms & Conditions do not prohibit any specific type of content.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at: https://www.livegore.com/terms-and-conditions
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	No specific procedures.

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)	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?	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. Main changes since last Report	LiveGore.com was not included in the last Report.

10. ChirpWire

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. ChirpWire's Terms of Use (ChirpWire, 2023) state that users must agree to use the services only to send and receive messages and material that are proper and related

	to the particular service, area, group, forum, community or other message or communication facility. In addition to any other terms or conditions of use of any bulletin board services, chat areas, news groups, forums, communities and/or other message or communication facilities, users should not: • Publish, post, upload, distribute or disseminate any inappropriate, profane, derogatory, defamatory, infringing, improper, obscene, indecent or unlawful topic, name, material or information • Defame, abuse, harass, stalk, threaten or otherwise violate the legal rights (such as rights of privacy and publicity) of others
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://chirpwire.net/terms_of_use
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?	ChirpWire states it has no obligation to monitor the bulletin board services, chat areas, news groups, forums, communities and/or other message or communication facilities. However, ChirpWire reserves the right at all times to disclose any information deemed by ChirpWire necessary to satisfy any applicable law, regulation, legal process or governmental request, or to edit, refuse to post or to remove any information or materials, in whole or in part. ChirpWire may share a user's personal information with third parties if it believes that the sites and services are being used in the commission of a crime, including to report such criminal activity or to exchange information with other companies and organisations for the purposes of fraud protection and risk management.
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)	ChirpWire does not control or endorse the content, messages or information found in any bulletin board services, chat areas, news groups, forums, communities and/or other message or communication facilities and, specifically disclaims any liability with regard to same and any actions resulting from your participation.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	See section 4 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.
11. Main changes since last Report	ChirpWire was not included in previous Reports.

11. 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 Terms of Use (Element, 2023) specifically state that users must not upload any material that could incite a terrorist offence, solicit any person to participate in terrorist activities, provide instruction on any method or technique for committing a terrorist offence or threaten to commit a terrorist offence. Element's Terms of Use prohibit the use of its platform: • To upload terrorist content • In any way that breaches any applicable local, national or international law or regulation. • In any way that is unlawful or fraudulent or has any unlawful or fraudulent purpose or effect. • For the purpose of harming or attempting to harm minors in any way. • To bully, insult, intimidate or humiliate any person.

	• To send, knowingly receive, upload, download, use or re-use any material which does not comply with the Content standards. In addition, Element's content standards state that contributions from users must not: • Be obscene, offensive, hateful or inflammatory • Bully, insult, intimidate or humiliate • Promote violence • Promote discrimination based on race, ethnicity, national origin, caste, sexual orientation, • gender, gender identity, religious affiliation, age, disability or disease • Promote any illegal content or activity • Encourage behaviour prejudicial to health or safety
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://element.io/legal
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?	Element states that whenever a user makes use of a feature that allows to upload content to its platform, or to make contact with other users of its platform, they must comply with the sections headed Prohibited uses and Content standards. Element has the right to remove any posting a user makes on the platform if, in Element's opinion, their post does not comply with the Terms of Use.
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)	If a user becomes aware of any material hosted on Element servers that could comprise or be connected to child sexual abuse or exploitation or that could comprise terrorist content or be connected to terrorism, they may contact Element on abuse@element.io. Element adds that it has limited ability to moderate hosted cloud customers

	content (it cannot amend or delete content or suspend users) as Element is generally not provided with access to such content by the customer. If Element receives any such complaint about content, it shall contact the customer with the details provided, and this shall be the extent of the action Element can be required to take. If a user wishes to complain about any other content not hosted on Element servers, including selfhosted customers, they should also contact Element on abuse@element.io. Likewise, Element adds that it has limited ability to moderate hosted cloud customers content (it cannot amend or delete content or suspend users) as Element is generally not provided with access to such content by the customer. If Element receives any such complaint about content, it shall contact the owner of the homeserver with the details provided, and this shall be the extent of the action Element can be required to take. In some circumstances it may be able to (but shall not be obliged to) render certain homeservers as unsearchable or apply a “not safe for work” or “NSFW” filter against said homeserver or room. Element states that its platform may include information and materials uploaded by other users of the platform, including to social media pages, video-sharing platforms, bulletin boards and chat rooms. This information and these materials have not been verified or approved by Element. The views expressed by other users on the platform do not represent Element’s views or values. Lastly, Element explains that where its platform links to other sites or platforms and resources provided by third parties, these links are provided for the users’ information only. Such links should not be interpreted as approval by Element of those linked sites or platforms or information users may obtain from them. Element adds that it has no control over the content of those sites, platforms or resources.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	When Element considers that a breach of its Terms of Use has occurred, it may take such action as it deems appropriate. Any breach may result in Element taking all or any of the following actions: • Immediate, temporary or permanent withdrawal of the user’s right to use the platform

	• Immediate, temporary or permanent removal of any Contribution uploaded by the user to the platform • Issue of a warning to the user • Legal proceedings against the user for reimbursement of all costs on an indemnity basis (including, but not limited to, reasonable administrative and legal costs) resulting from the breach • Further legal action against the user • Disclosure of such information to law enforcement authorities as Element reasonably feels is necessary or as required by law Element excludes its liability for all action it may take in response to breaches of its Terms of Use. The actions it may take are not limited to those described above, and it may take any other action it reasonably deems 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. Main changes since last Report	• Element provides a more precise list of the different actions it may take in case of breaches of its Terms of Use.

12. Gab

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 (Gab, 2023) 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	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 believes 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 judgement 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 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.
11. Main changes since last Report	No main changes since last Report.

13. Mastodon

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	Mastodon does not specifically prohibit TVEC. Mastodon broadly provides the following “Server rules”: • No incitement of violence or promotion of violent ideologies • No racism, sexism, homophobia, transphobia, xenophobia, or casteism • Sexually explicit or violent media must be marked as sensitive when posting (Mastodon, s.d.)
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://mastodon.social/about
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 6 below.
4.1 Notifications of removals or other enforcement decisions	All default user moderation decisions will notify the affected user by email.
4.2 Appeal processes against removals or other enforcement decisions	The user can access an appeal page, where they can submit one appeal within 20 days of the decision. Moderators can approve or reject the appeal.
5. Means of identifying TVEC (for example, monitoring algorithms, user-generated, human (staff) reviewers, hash-sharing/URL sharing database)	Users can report problematic content to moderators using the “report” option, adding a note about why they are reporting the account. Users can attach certain problematic statuses for additional context on why they are reporting the account, and if their conduct is violating the rules of the remote website, they can also choose to forward the report to their site’s moderators (Mastodon, 2023). When reporting accounts, Mastodon encourages users to include at least a few posts that show rule-breaking behaviour, when applicable. If there is any additional context that might help make a decision, users should include it in the comment. This is especially important

	when the content is in a language nobody on the moderation team speaks. Mastodon has a team of paid moderators that usually handles reports within 24 hours. Users are not notified when a report they have made has led to a punitive action. Mastodon adds that not all punitive actions are externally visible (Mastodon, s.d.). In 2022, Mastodon added the ability to quickly suspend all accounts matching specific search queries, such as a matching IP range or email domain. Mastodon also added a webhook system to allow server operators to setup more elaborate automation systems for their moderation needs, and expanded the set of moderation APIs to support it (Mastodon, 2023).

6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	On Mastodon, users can control what they see for a more comfortable social media experience. The following tools are at their disposal:
	
	• Filtering posts
	  ○ Keyword or phrase
	  ○ Expire after
	  ○ Filter contexts
	  ○ Drop instead of hide
	  ○ Whole word
	• User-level actions
	  ○ Hiding boosts
	  ○ Muting
	  ○ Blocking
	  ○ Hiding an entire server
	
	In addition, moderators can take a range of action against unwanted users or domains (Mastodon.social, 2023):
	
	• Moderating individual users
	  ○ Sensitive
	  ○ Freeze
	  ○ Limit
	  ○ Suspend
	• Moderating entire websites
	  ○ Reject media
	  ○ Limit
	  ○ Suspend
	
	Lastly, Mastodon provides a list of all limited or suspended servers with the corresponding reason, among which are

	"hate speech" and "inappropriate content" (Mastodon, s.d.).
7. Does the service issue transparency reports (TRs) on TVEC?	No. Mastodon publishes annual reports with general information on their services including content moderation. However, they do not feature transparency reporting on TVEC (Mastodon, 2023).
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. Main changes since last report	Mastodon was not included in previous Reports.

14. TamTam.Chat

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, TamTam's "Chats and Channels Regulations" (TamTam.Chat, 2022) state that messages in channels and chats must not: • Promote and call for violence and cruelty, committing suicide and other illegal and immoral acts. • Propagate extremism, terrorism, excite hostility based on racial, ethnical or national identity, sexual orientation, gender, gender identity, religious opinions, age, limited physical or mental abilities or diseases. • Contain images and video of low quality, obscene, pornographic, images of dead people and animals, with marks of violence, cruelty and other scaring or aesthetically unacceptable images. • Contain other information of an illegal nature.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://about.tamtam.chat/en/regulations/

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?	TamTam states that if there are any violations related to a public channel or public chat name, published materials, channel maintenance or promotion process, the public channel or public chat may be blocked permanently, without a chance to be unblocked.
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 via the link tt.me/support or by mailing to abuse@tamtam.chat.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Channels and public chats that do not comply with TamTam's Chat and Channels Regulations may be blocked by the decision of the Administration.
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. Main changes since last report	TamTam.Chat was not included in previous Reports.

15. Matrix

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	Matrix does not specifically prohibit TVEC. Matrix's Code of Conduct (Matrix, 2023)prohibits harassment, which includes but is not limited to: • 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 • 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 In addition, the Matrix.org Homeserver Terms and Conditions prohibit to use the Service for any unlawful purposes or in support of illegal activities under UK/EU law. By using the Service, users agree to comply with all applicable laws governing their online conduct and content.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://matrix.org/legal/terms-and-conditions/ and 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?	Matrix's Terms and Conditions state that any form of illegal content is strictly prohibited in the Service and its distribution will result in immediate account termination.
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 experiencing or witnessing unacceptable behaviour (or have any other concerns) can report via abuse@matrix.org. Matrix states that all reports are handled with discretion and should include: • The user's contact information. • Names (usernames and nicks, real names, and/or pseudonyms) of any individuals involved. If there are additional witnesses, the user should include them as well. • An account of what occurred, and if the user believes the incident is ongoing. • The date and time of the incident (or start of incident). • Any additional information that may be helpful. After filing a report, the user will receive an automated confirmation email. Typically, Matrix does not answer abuse reports unless further clarification is required. Moreover, Matrix uses Mjolnir, an all-in-one automated tool to handle content moderation of rooms under its Code of Conduct (Matrix, 2023).
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Matrix has a basic system of roles, sometimes also called Power Levels (usually a number going from 0 to 100). By default, there are three roles in Matrix: • (0) Users can send messages, media, reactions and redact their own messages in a room • (50) Moderators can also change the room name, address, topic, remove users from the room (temporarily or permanently), redact other's messages and send a notification to everyone in the room • 100: Administrators can also change the history visibility (whether people can see messages from before they joined or not), enable encryption in the room, remove entire servers from the room, and promote others to Moderator or Administrator. Content moderation tools available to moderators and administrators are as follows (Matrix, 2023): • Redacting a specific message • Redacting a user's last messages • Removing someone temporarily (kick) • Removing someone definitely (ban) • Removing a server definitely

	• Creating ban lists • Subscribing to ban lists Moreover, Matrix's automated content moderation tool (called Mjolnir) by default includes support for bans, redactions, anti-spam, server ACLs, room directory changes, room alias transfers, account deactivation, room shutdown, and more.
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. Main changes since last report	Matrix was not included in previous Reports.

16. Abolitionmedia.noblogs.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 available.
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	Not applicable.

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	No information available.
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. Main changes since last report	Abolitionmedia.noblogs.org was not included previous Reports.

17. Americanfuturistpublishing.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 available.
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)	No information available.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	No information available.
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. Main changes since last report	Americanfuturistpublishing.com was not included previous Reports.

18. Malhm.xyz

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 available.
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)	No information available.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	No information available.
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. Main changes since last report	Malhm.xyz was not included previous Reports.

19. Alazaimll.websites.co.in

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 available.
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)	No information available.

6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	No information available.
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. Main changes since last report	Alazaimll.websites.co.in was not included previous Reports.

20. Umarmediattp.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 available.
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)	No information available.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	No information available.
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. Main changes since last report	Umarmediattp.org was not included previous Reports.

21. Shahadanews.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 available.
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	Not applicable.

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	No information available.
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. Main changes since last report	Shahadanews.com was not included previous Reports.

22. 4chan

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, 4chan's ToS (4chan, 2023) prohibit users from uploading, posting, discussing, requesting, or linking to anything that violates local or United States laws.

2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://www.4chan.org/rules
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	According to 4chan, threads expire and are pruned by 4chan's software at a relatively fast rate. Since most boards are limited to ten pages, content is usually available for only a few hours or days before it is removed. Usually, missing posts were probably pruned automatically; however, in some cases they may have been removed by a moderator or 'janitor'. Moderators are individuals selected to perform general site maintenance. They may delete posts globally, ban users, close threads and carry out associated actions. Janitors are a class between 'end user' and 'moderator'. They are given access to 4chan's report system and may delete posts on their assigned board(s), as well as submit ban requests. Janitors are selected via an application, orientation, and testing process. Admission to the moderation team is by invitation only. The janitor program is occasionally opened to new applicants. There is no public record of content deletion and because threads are frequently pruned, there is no way of knowing which pieces of content have been removed by the moderation team. In short, there is no way for an end user to any given point in time. The 4chan moderation team reserves the right to block or ban access and remove content for any reason without notice. Users are temporarily blocked from posting when there is a pending ban request placed on their IP address. This block lasts 15 minutes from the time a janitor submits a ban request and is removed immediately if the request is denied by a moderator. If the request is approved, a regular ban is applied.
4.1 Notifications of removals or other enforcement decisions	No notifications are specified.

4.2 Appeal processes against removals or other enforcement decisions	Users can appeal bans it they believe an error has been made, by contacting the moderators.
5. Means of identifying TVEC (for example, monitoring algorithms, user-generated, human (staff) reviewers, hash-sharing/URL sharing database)	4chan states that it encourages reporting posts for review. Moderators review the reported content and take appropriate action.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Breaking 4chan's Rules may result in post deletion, a temporary ban, or in some cases, permanent banishment.
7. Does the service issue transparency reports (TRs) on TVEC?	No.
8. What information/fields of data are included in the TRs?	Not applicable.
9. Methodologies for determining/calculating/estimating the information/data included in the TRs	Not applicable.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Main changes since last report	4chan was included in the first edition of this Report (2020) as part as the global top 50 online content-sharing services. There were no main changes to 4chan's profile since then.

23. Nuceciwan127.xyz

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 available.
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)	No information available.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	No information available.
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. Main changes since last report	Nuceciwan127.xyz was not included previous Reports.

24. Amjaad.video

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 available.

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)	No information available.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	No information available.
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. Main changes since last report	Amjaad.video was not included previous Reports.

25. Dalelansar.info

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 available.
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	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	No information available.
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. Main changes since last report	Dalelansar.info was not included in previous Reports.

26. Ansarollah.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 available.
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	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	No information available.
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. Main changes since last report	Ansarollah.com was not included in previous Reports.

27. Alqassam.ps

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 available.
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	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	No information available.
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. Main changes since last Report	Alqassam.ps was not included in previous Reports.

28. Threads

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	Threads' ToS supplement and amend Instagram's ToS and Community Guidelines (Instagram, 2023). Instagram and Facebook share content policies, so their ToS and Guidelines are incorporated by reference. Facebook notes that if content is considered to be in violation of such policies on Facebook, it is also considered violating on Instagram, therefore also on Thread. 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. Under a specific section titled 'Dissemination of Terrorist Content Online', Instagram refers to the Regulation Addressing Dissemination of Terrorist Content Online (EU 2021/784) (TCO) that provides information for EU Member State competent authorities on how to report terrorist content (Instagram, 2023). This Article is only relevant to people in the EU. According to TCO, reasons for considering material to be terrorist content are that such material: • Incites others to commit terrorist offences, such as by glorifying terrorist acts, by advocating the commission of such offences; • Solicits others to commit or to contribute to the commission of terrorist offences; • Provides instruction on the making or use of explosives, firearms or other weapons or noxious or hazardous substances, or on other specific methods or techniques for the purpose of committing or contributing to the commission of terrorist offences; or • Constitutes a threat to commit one of the terrorist offences. More information on terrorist offences can be found in Article 3(1) of Directive (EU) 2017/541 (available in the Official Journal of the European Union). Instagram maintains a separate Community Guidelines that apply worldwide. A section ('Follow the law') in the Guideline states that Instagram is not a place to support or praise 'terrorism, organised crime or hate groups' (Instagram, 2023), which contains the link that directs one to the Transparency Centre of its operating company, Meta. The same Community Standards on 'Dangerous Organisations and Individuals' apply to Threads/Instagram (Meta, 2023). See a section on Facebook in Annex B. Moreover, Instagram removes content that contains 'credible threats' or 'hate speech, and prohibits 'serious threats of harm' to public and personal safety, all of which contains the link directed to Meta's Transparency Centre. According to Meta's policies, those behaviours are categorised under the 'Violence and Incitement' policy. See a section on Facebook in Annex B.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Threads/Instagram's Community Guidelines are available at https://www.facebook.com/help/instagram/477434105621119/ Threads/Instagram's ToS are available at https://help.instagram.com/581066165581870

3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	Meta has the right to: • remove any Threads content that is stored on Threads servers from being accessible within the Threads service if Meta believes that such contents violate their Terms or that of Instagram's or if Meta is permitted or required to do so by applicable law; • provide notice to Third Party Servers about enforcement actions taken with respect to Threads content; and • take any action that meta believes is necessary or appropriate if meta reasonably believes that any such Threads content infringes the rights of others and/or could create liability or adverse legal or regulatory implications for Meta or other Threads users. Instagram removes or blocks any content or information users share if they believe that it violates the ToS, their policies, or if they are required to do so by law. In some cases when contents are removed, Instagram may let the users know and explain any options they have to request another review, unless one seriously or repeatedly violates the ToS or if doing so may pose the company or others to legal liability. Thread/Instagram applies the same Complaint Handling Process established by Meta (Meta, 2023) and invites the users to consult their Help Center (Instagram, 2023).
4.1 Notifications of removals or other enforcement decisions	If content goes against Instagram's Terms of Service or Community Guidelines, Threads/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	Threads has the same appeal process as Instagram and Facebook. See section 4.2 of Facebook's profile.
5. Means of identifying TVEC (for example, monitoring algorithms, user-generated, human (staff)	Threads uses the same methods as Instagram and Facebook. See Section 5 of Facebook's and Instagram's profiles.

reviewers, hash-sharing/URL sharing database)	
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Threads can refuse to provide or stop providing all or part of the services. This includes terminating or disabling the profiles and access to the Threads service without notice or after providing reasonable notice where required by applicable law. They can also do so in their reasonable discretion (Instagram, 2023).
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. Main changes since last Report	Threads was not included in previous Reports.

29. GoyimTV.com

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 category 'Respect and Decency', GoyimTV's ToS state that terrorist material, credible threats or incitement to violence, and malicious use of the platform will not be tolerated.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at: https://goyimtv.com/page?t=page-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?	No specific procedures. Its ToS state that GoyimTV reserves the right to remove any of the users' content at any time without 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)	No information available.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	See section 4 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.
11. Main changes since last Report	GoyimTV.com was not included in previous Reports.

30. Moqawama.ir

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 available.
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?	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?	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	No information available.
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. Main changes since last Report	Moqawama.ir was not included in previous Reports.

31. 3pdirectory.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 available.
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	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,	No information available.

user-generated, human (staff) reviewers, hash-sharing/URL sharing database)	
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	No information available.
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. Main changes since last Report	3pdirectory.com was not included in previous Reports.

32. Odysee

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 (Odysee, 2022) 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), 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;

	• 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	Available at https://help.odysee.tv/communityguidelines/ and 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 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).
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.
11. Main changes since last Report	No main changes since last Report.

33. YouTube

See profile 2 in Annex B.

34. Archive.org (Internet Archive)

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, Archive.org's ToS provide that users may not 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	Available at https://archive.org/about/terms.php
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.

4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	No procedures are specified. The Archive.org broadly states that if a user does not want one's work to be in their collections, they may remove that portion without 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)	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.
11. Main changes since last Report	No main changes since last Report.

35. Justpaste.it

1. How is terrorist and violent extremist content (TVEC) defined in the Terms of Service (ToS) or Community Guidelines/Standards?	The ToS define 'terrorist content' by referring to EU and UN definitions. "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". In addition, JustPaste.it is committed to respecting the internationally recognised human rights set out in the United Nations Guiding Principles on Business and Human Rights (UNGPs), especially to protecting and supporting freedom of speech and privacy. Justpaste.it explains that it puts its users first and want to ensure that everyone feels safe and welcome as a member of its community as long as their activity is compliant with its ToS.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at: https://justpaste.it/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?	JustPaste.it reserves the right to decide the compliance of content with the requirements set out in ToS and my remove content and/or terminate the user accounts if in violation of the requirements without prior notice and at its sole discretion.
4.1 Notifications of removals or other enforcement decisions	Contents and/or accounts may be removed and/or terminated without prior notice.
4.2 Appeal processes against removals or other enforcement decisions	Users may appeal by sending an email to support@justpaste.it (JustPaste.it, 2023).
5. Means of identifying TVEC (for example, monitoring algorithms, user-generated, human (staff)	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

reviewers, hash-sharing/URL sharing database)	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).
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Contents and/or accounts may be removed and/or terminated without prior notice
7. Does the service issue transparency reports (TRs) on TVEC?	Yes.

8. What information/fields of data are included in the TRs?	In its last TR (covering 2022) (Justpaste.it, 2022), Justpaste.it reported the following information:
	• Total number of requests, broken down by country
	  ○ EU
	  ○ UK
	  ○ Türkiye
	• Percentage of requests in relation to terrorist materials
	• Percentage of such requests in relation to terrorist materials in which Justpaste.it took action and blocked the 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.
11. Main changes since last Report	• JustPaste.it is committed to respecting the internationally recognised human rights set out in the United Nations Guiding Principles on Business and Human Rights (UNGPs).

36. Google Drive

See profile 46 in Annex B.

37. SoundCloud

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, under its "Terrorist content" policy, SoundCloud's Community Guidelines (SoundCloud, 2023) state that SoundCloud strictly prohibits the dissemination of terrorist content and is committed to its prompt removal from the platform. While Soundcloud takes into consideration UN, US and EU lists of terrorist designated organisations, it also follows and enforces the details Pursuant to Article 3 of Regulation (EU) 2021/784 of the European Parliament and of the Council. Content allowed under this policy: • Speak about terrorist groups in a purely educational manner or for reporting through legitimate news sources, making it clear when the user is quoting directly and not presenting them as their own views. • Specifically state that the user does not endorse the groups/events/etc. Content prohibited under this policy: • Upload content glorifying terrorist acts, groups or leaders • Comment in support of, or encouraging a terrorist attack In addition, under its "Violence and threatening behaviour" policy, SoundCloud does not allow any use of intimidation or threats of violence between its users, against the wider public or against SoundCloud employees. Likewise, SoundCloud does not permit any incitement to, or glorification of, violent acts either against an individual or the wider public. Additionally, the use of excessive gore in regard to both imagery and language is not permitted. Content allowed under this policy: • Provide commentary on violence in regard to a specific event (i.e., something that happened to the user in the context of a podcast episode).

	• Refer to violence where this serves an educational purpose. Content prohibited under this policy: • Upload content that has the intention to intimidate an individual into fearing a physical attack against their person. • Upload content inciting violence (i.e., calling on or encouraging others to complete violent acts against an individual or against the wider public). • Upload graphic images of real or dramatized violence with the intent to glorify violent acts or to shock/disgust. Lastly, under its "Hate speech" policy, Soundcloud does not allow behaviour or content that promotes or encourages hatred, discrimination or violence against others based on race, cultural identity or ethnic background, religious beliefs, disability, gender identity, or sexual orientation. SoundCloud will use its discretion when necessary to determine if beliefs or opinions incite hatred. Users should keep in mind that applicable laws state that Nazi and ISIS content is illegal, and will be removed without discussion, unless clearly used for educational or journalistic purposes. Content allowed under this policy: • Speak about a hate group in a purely educational manner, making it clear when quoting directly and not presenting them as the user's own views. • Specifically state that the user does not endorse the groups/events/etc. that they may be quoting. Content prohibited under this policy: • "I mean, these [protected group], they know what's wrong with them" • "[protected group] are threatening us. We need to get them out"
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://soundcloud.com/community-guidelines
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	Not applicable.

4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	If SoundCloud believes that a user's activity violates its guidelines, Terms of Use, any other associated policies, or applicable law, it will send a written warning to the email associated with the account and a statement of reasons. After more than two of these warnings, SoundCloud may suspend or terminate the user's account at its discretion. In certain cases, such as when SoundCloud deems an account to be dedicated to the violating activity, it reserves the right to terminate the account immediately, without prior warning, at its discretion. On suspension or termination of an account, the user will be notified with a statement of reasons and given the possibility to dispute.
4.1 Notifications of removals or other enforcement decisions	SoundCloud notifies users of removals and other enforcement decisions via email.
4.2 Appeal processes against removals or other enforcement decisions	Where content is removed or a suspension/termination is issued, users have the right to appeal the decision made in all instances (subject to the limits described below). SoundCloud clearly sets out the rights of all parties to appeal a decision and how to do so in the written warning or notice sent to the user during the complaints procedure, as well as within the applicable notification. Appeals are routed in a manner that ensures impartiality, meaning that no appeal is reviewed by the same party that made the original decision. Where a decision on an appeal is not forthcoming, there is a clear line of escalation to ensure a renewed decision is available. Users may file an appeal at any time following an initial decision. If the appeal is unsuccessful, SoundCloud will explain to the user the reasons why. If the appeal is successful, the content will be restored. Users may also have additional appeal options, such as the possibility of an out-of-court dispute settlement process or to take legal action in court. Where required by applicable laws, some adjustments to the appeals procedure may apply, to ensure that SoundCloud is complying with its legal obligations (SoundCloud, 2024).
5. Means of identifying TVEC (for example, monitoring algorithms, user-generated, human (staff) reviewers, hash-sharing/URL sharing database)	SoundCloud explains that it has the right, but not the obligation, to review and moderate the content on the Platform and to take actions against content that is illegal or otherwise objectionable. In the following policy, "moderate" means taking action, either by automation or not, to identify or address illegal content, content that breaches the Terms, or otherwise

	objectionable activity. For clarity, SoundCloud does not actively monitor content or the Platform. SoundCloud employs numerous automated processes and techniques to address illegal and objectionable content on its Platform. At the point of sign up, users are required to agree to the Terms of Use and Community Guidelines that clearly state that users are not permitted to upload any illegal content or content that breaches those documents. Prior to the upload of content or direct interaction with other users of SoundCloud, users must verify their email address, which is a measure aimed at ensuring only genuine actors utilise SoundCloud, mitigating the risk of the anonymous proliferation of violative content. <u>Content moderation</u> At its discretion, SoundCloud may use automated and manual tools to detect and prevent objectionable content. This is currently performed via two approaches: • Detection of violative profile names, URLs, and profile descriptions: Words or phrases found to be a breach of SoundCloud's ToS will be prohibited from use in profile names, profile descriptions, and URLs. Any attempted use of violative words or phrases may result in automatic and permanent suspension from the platform. • Detection of violative audio, images, and track data: A third party vendor is used to proactively detect and remove violative uploaded audio, images and track data content. Image detection, audio matching and text recognition among other techniques are utilised in the detection of violative content. Detected content is either flagged for review or automatically suspended. Where content is flagged and subsequently removed, the account holder is issued a notification including a statement of reasons via the primary email associated with the account. Content removed in this manner will contribute towards a user's status as a Repeat Infringer (see Section 6 below). <u>Complaint procedure and decision-making process</u> SoundCloud operates a report and takedown process (complaints procedure) for content suspected to breach its Terms or the law. Users are encouraged to report any

	instances of violative content or behaviour on the platform. Depending on location, the complaints procedure can be initiated either via dedicated report forms in SoundCloud's help centre or report buttons. All complaints are routed to SoundCloud's Trust & Safety Team for review. SoundCloud states that its reporting options are easy to find and use, and the complaints procedure is streamlined and straightforward. To streamline the reporting process and make sure it reaches the relevant team as quickly as possible the notification form contains drop-down menus with different categories of violations for users to choose from. Notices must contain: • sufficiently substantiated explanation of the reasons why the user alleges the information to be violative, • clear indication of the URL, • the user's name and email address, and • a confirmation that the information and allegations it contains are accurate and complete. Insufficient notices will be rejected. Upon lodging a complaint, a confirmation of receipt will be sent to the email address provided in the complaint (SoundCloud, 2024).
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	• Content restrictions: SoundCloud reserves the right to block, remove, delete, limit or restrict access to any content at any time, without liability, including without limitation, if it has reason to believe that such content does or might infringe the rights of any third party, has been uploaded or posted in breach of the Terms or law, or is otherwise unacceptable to SoundCloud. If SoundCloud removes or restricts access to content, where required by applicable laws it will notify the user and explain the reason for the decision. • Repeat infringers – Protection against misuse SoundCloud may suspend or terminate access to the Platform if SoundCloud determines, in its reasonable discretion, that a user has repeatedly breached the Terms, for example where their content repeatedly breaches the Terms of Use or frequently submit notices or complaints (or appeals) that are manifestly unfounded ("Repeat Infringers"). When considering if a user has breached the Terms, and to determine the steps that SoundCloud will take as a result of that behaviour such as duration of suspension, SoundCloud will take into account factors and circumstances such as the nature, number,

	seriousness and gravity of any such violations, and all types of content uploaded in breach of the Terms. For example, if a user repeatedly reports content as violative and SoundCloud finds, repeatedly that the claims are unfounded, SoundCloud may suspend the account for a period of [x] days. • Termination SoundCloud may suspend or terminate a user's access to the Platform (or certain features of the Platform) at any time if: (i) they are deemed to be a Repeat Infringer as described above; or (ii) they are in breach the Terms. Finally, SoundCloud may also suspend or terminate an account without warning if ordered to do so by a court, where applicable laws require SoundCloud to do so, and/or in other appropriate circumstances, as determined by SoundCloud at its discretion.
7. Does the service issue transparency reports (TRs) on TVEC?	No. SoundCloud issues bi-annual reports under German law (German Network Enforcement Act or NetzDG).
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. Main changes since last Report	SoundCloud was not included in previous Reports.

38. Dropbox

See profile 47 in Annex B.

39. 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	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?	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?	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 Filtures

	- 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.
11. Main changes since last Report	No main changes since last Report.

40. 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 available.
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?	No.
4. Policies and procedures to implement and enforce the ToS or Community	Not applicable.

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)	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?	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. Main changes since last Report	No main changes since last Report.

41. Itarchives.org

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.
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?	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?	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)	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?	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. Main changes since last Report	Itcarchive.com was not included in the last Report.

42. DoxBin.org

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, DoxBin's ToS state that any support of terrorism or threats of physical violence is not allowed in its introductory section titled 'What is Doxbin?'.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at: https://doxbin.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 specific policies and procedures. However, ToS states that the posts that do not comply with their list of rules will be 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)	Users wishing to request for a paste to be removed must contact Doxbin's staff on Telegram at @Brenton or @Doxer.

reviewers, hash-sharing/URL sharing database)	
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	Posts may 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.
11. Main changes since last Report	No main changes since last Report.

43. File.io

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. File.io's ToS prohibit content that is hate speech, threatening or pornographic, that incites violence or that contains nudity or graphic or gratuitous violence.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at: https://www.file.io/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	File.io states that it may review Third Party Content to determine whether it is illegal or violates its policies, and it may remove or refuse to display Third Party Content that it believes violates its policies or the law. However, File.io does not

content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	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	In case of breach of its ToS, File.io may suspend or stop the provision of its services to the infringer, as well as remove the 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.
11. Main changes since last Report	No main changes since last Report.

44. 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' which refers to 'videos, images or audio fragments which promote and glorify acts of terrorism'. Similarly, gory contents are not allowed, that are, 'graphic and shocking videos or images depicting severe harm to humans (or animals)'.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at: https://pixeldrain.com/abuse
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 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	Content may 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.
11. Main changes since last Report	No main changes since last Report.

45. Gofile.io

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. 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	Available at: https://gofile.io/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?	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)	Not applicable.
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.
11. Main changes since last Report	No main changes since last Report.

46. Signal

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. Signal's ToS broadly state that users should use its services only for legal, authorised, and acceptable purposes. Users should not use Signal in ways that involve sending illegal or impermissible communications. In addition, Signal explains that when users use third-party services, those third-party services' terms and privacy policies apply.

2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at: https://signal.org/legal/#terms-of-service
3. Are there specific provisions applicable to livestreamed content in the ToS or Community Guidelines/Standards?	No.
4. Policies and procedures to implement and enforce the ToS or Community Guidelines/Standards (removal of content). In particular: are there notifications of removals or other enforcement decisions and appeal processes against them?	Signal may modify, suspend, or terminate a user's access to or use of its Services anytime for any reason, such as if they violate the letter or spirit of the Terms or create harm, risk, or possible legal exposure for Signal. If Signal disables a user's account for a violation of its Terms, they should not create another account without permission.
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)	Not applicable.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	See Section 4 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.
11. Main changes since last Report	Signal was not included in previous Reports.

47. Wire

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. Wire's ToS prohibit content that is "unlawful or illegal, defamatory, harmful, abusive, hateful, racially or ethnically offensive that encourages conduct that would be considered a criminal offence".
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at: https://start.wire.com/en-us/terms-of-use-personal for personal users; and https://start.wire.com/en-us/en-us/terms-of-use-personal-1 for business users.
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?	Wire may disable a user's account if they determine that the user is: • in breach of or otherwise acting inconsistently with the ToS; or • engaging in fraudulent or illegal activities or other conduct that may result in liability to the 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)	See Section 4 above.

6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	In case of breach of ToS, Wire may disable the user's account.
7. Does the service issue transparency reports (TRs) on TVEC?	No. The service issues TRs, however, not specifically on TVEC (Wire, 2023).
8. What information/fields of data are included in the TRs?	TRs include details about how often authorities request user data from their service. However, only an overview is made publicly available, for instance, on the number of formal requests made to their service.
9. Methodologies for determining/ calculating/estimating the information/data included in the TRs	No.
10. Frequency/timing with which TRs are issued	Not applicable.
11. Main changes since last Report	Wire was not included in previous Reports.

48. Slack

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. According to Slack's Acceptable Use Policy (Slack, 2021) users must not: • use the Services to provide material support or resources (or to conceal or disguise the nature, location, source, or ownership of material support or resources) to any organisation(s) designated by the United States government as a foreign terrorist organisation pursuant to section 219 of the Immigration and Nationality Act or other laws and regulations concerning national security, defence or terrorism; • engage in activity that incites or encourages violence or hatred against individuals or groups;

	• authorise, permit, enable, induce or encourage any third party to do any of the above. In addition, Slack states that users must comply with all applicable laws and governmental regulations; and promptly notify Slack if they become aware of or reasonably suspect any illegal or unauthorized activity or a security breach. Lastly, Slack's Community Forum Terms of Service (Slack, 2021) prohibit users from: • posting content that is abusive, offensive, vulgar, obscene, hateful, racist or bigoted, threatening, libelous, defamatory, or fraudulent; • posting content that violates applicable laws and governmental regulations.
2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://slack.com/acceptable-use-policy; and https://slack.com/terms-of-service/user
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 Slack believes that there is a violation of the Contract, User Terms, the Acceptable Use Policy, or any of Slack's other policies that can simply be remedied by Customer's removal of certain Customer Data or taking other action, Slack will, in most cases, ask Customer to take action rather than intervene. Slack may directly step in and take what it determines to be appropriate action (including disabling the user's account) if Customer does not take appropriate action or if Slack believes there is a credible risk of harm to anyone.
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)	If Administrators have enabled content flagging on a given Slack Workspace, users can flag messages for review. Slack does not review flagged content. The flagged messages are sent to designated administrators for review. Content Administrators are designated by the Workspace Primary Owners. These Administrators can see the name of

	the user making the request, the flagged content, and the reason for flagging the content. Other users will not be notified.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	If Slack believes that a violation of the policy is deliberate, repeated or presents a credible risk of harm to other users, its customers, the Services or any third parties, Slack may suspend or terminate the user's access (Slack, 2021). When content is flagged, Content Administrators can decide to dismiss the report, or hide the content from a Workspace (Slack, 2024).
7. Does the service issue transparency reports (TRs) on TVEC?	No. Slack issues transparency reports with information about government requests for user data but 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. Main changes since last Report	Slack was not included in previous Reports.

49. WhatsApp

See profile 4 in Annex B.

50. Threema

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 category 'Provisions of Use', it is stated that 'unauthorised content' broadly includes the content 'prohibited by applicable criminal laws' without more specific examples (Theema, 2022).

2. Manner in which the ToS or Community Guidelines/Standards are communicated	Available at https://threema.ch/tos/?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?	Threema reserves the right to suspend a User's Authorisation, restrict the availability of the Service or revoke a User's Authorisation without consulting the User or stating reasons, provided that Threema suspects that the User has breached the Terms of Use or committed unlawful acts.
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)	If a user finds communication of unauthorised content, the user may report it via email to abuse@threema.ch.
6. Sanctions/consequences in case of breaches of the ToS or Community Guidelines/Standards	No information provided.
7. Does the service issue transparency reports (TRs) on TVEC?	No. Threema issues TRs, however, not specifically on TVEC (Threema, 2023).
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. Main changes since last Report	Threema was not included in previous Reports.

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

4chan. (2023). *Global Rules*. Retrieved from https://www.4chan.org/rules

Access Now. (2020). *Two Years under the EU GDPR.*

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

Alexander, S. (2023, September). Retrieved from https://china-digital.com/blogs/chinas-social-networks/#:~:text=Baidu%20Tieba%20social%20network&text=300%20million%20monthly%20active%20users%20(MAU)

Ali, U. (2023, September 13). *Umar Media: The TTP's Media Wing*. Retrieved from https://www.linkedin.com/pulse/umar-media-ttps-wing-usman-ali/

Anti-Defamation League. (2020, April 5). *What is "Zoombombing" and Who is Behind It?* Retrieved from https://www.adl.org/resources/blog/what-zoombombing-and-who-behind-it

Anti-Defamation League. (2022, May 14). *Buffalo Shooter's Manifesto Promotes "Great Replacement" Theory, Antisemitism and Previous Mass Shooters*. Retrieved from https://www.adl.org/resources/blog/buffalo-shooters-manifesto-promotes-great-replacement-theory-antisemitism-and-previous-mass-shooters

Anti-Defamation League. (2023). *Goyim Defense League*. Retrieved from https://www.adl.org/resources/backgrounder/goyim-defense-league

Anti-Defamation League. (2023). *Hate on Display - Hate Symbols Database*. Retrieved from https://www.adl.org/resources/hate-symbols/search?keywords=&sort_by=title&page=3

Apple. (2022). *2022 App Store Transparency Report*. Retrieved from https://www.apple.com/legal/more-resources/docs/2022-App-Store-Transparency-Report.pdf

Apple. (2023). *ABout communication safety in Messages*. Retrieved from https://support.apple.com/en-us/HT212850

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

Article 19. (2023, August). *Content moderation and freedom of expression handbook*. Retrieved from https://www.article19.org/wp-content/uploads/2023/08/SM4P-Content-moderation-handbook-9-Aug-final.pdf

Automattic. (2023). *EU Terrorist Content Removal Orders*. Retrieved from https://transparency.automattic.com/wordpress-dot-com/eu-terrorist-content-removal-orders-2022-july-1-dec-31/

Automattic. (2023). *Wordpress.com Transparency Report*. Retrieved from Automattic: https://transparency.automattic.com/wordpress-dot-com/iru-reports/

Baidu. (2022). *Environmental, Social and Governance Report*. Retrieved from https://esg.baidu.com/ESG/Baidu_2022_ESG_Report.pdf

Baidu Tieba. (2023). *Tieba Agreement*. Retrieved from https://gsp0.baidu.com/5aAHeD3nKhI2p27j8IqW0jdnxx1xbK/tb/eula.html

Baillencourt, J. d. (2022, September 8). *Partnering to prevent violent extremism*. Retrieved from https://newsroom.tiktok.com/en-us/partnering-to-prevent-violent-extremism

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

Bateman, T. (2021, August 18). Retrieved from https://www.euronews.com/next/2021/08/18/taliban-whatsapp-accounts-highlight-social-media-companies-struggle-to-ban-afghanistan-gro

BBC. (2019). *Internet Archive denies hosting 'terrorist' content* .

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

Bilibili. (2021). *Bilibili 2021 Environmental, Social and Governance Report*. Retrieved from https://ir.bilibili.com/media/4nblhygw/bilibili-2021-environmental-social-and-governance-report.pdf

Bilibili. (2022). *Bilibili 2022 Environmental, Social and Governance Report*. Retrieved from https://ir.bilibili.com/media/agongs0v/bilibili-2022-environmental-social-and-governance-report.pdf

Bilibili. (2022, November 16). *Bilibili Community Rules*. Retrieved from https://www.bilibili.tv/marketing/protocal/communityrules_en.html

Bilibili. (2023). *Bilibili Black Room*. Retrieved from https://www.bilibili.com/blackroom/ban

Bilibili. (2023). *Discipline Committee*. Retrieved from https://www.bilibili.com/judgement/apply

Binder, J. F., & Kenyon, J. (2022). *Terrorism and the internet: How dangerous is online radicalization?* Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9606324/

Bleu, N. (2023, September). *21+ Top Quora Statistics For 2023 (Users + Growth)*. Retrieved from https://startupbonsai.com/quora-statistics/

Boyd, A. (2022, July). *How Xiaohongshu censors "sudden incidents"*. Retrieved from https://chinadigitaltimes.net/2022/07/how-xiaohongshu-censors-sudden-incidents/

Boyd, A. (2022, July 27). *How Xiaohongshu censors "sudden incidents"*. Retrieved from https://chinadigitaltimes.net/2022/07/how-xiaohongshu-censors-sudden-incidents/

Carbone, C. (2020). *FBI arrests alleged neo-Nazi linked to 'swatting'attacks on journalist*.

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

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

Chen, C. (2023, June 21). *China's banned online communities have found a new home on Reddit*. Retrieved from https://restofworld.org/2023/reddit-china-online-communities/

ChirpWire. (2023). *ChirpWire Terms of Use*. Retrieved from https://chirpwire.net/terms_of_use

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

Clegg, N. (2022, December 13). *Meta Launches New Content Moderation Tool as It Takes Chair of Counter-Terrorism NGO*. Retrieved from https://about.fb.com/news/2022/12/meta-launches-new-content-moderation-tool/

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

Counter Extremism Project. (2018, March 6). *Ads Found Alongside Bombmaking Video on Dailymotion*. Retrieved from https://www.counterextremism.com/blog/ads-found-alongside-bombmaking-video-dailymotion-0

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

Counter Extremism Project. (2020). *Extermist Content Online: ISIS Supporters Skirt Barriers on Telegram.*

Counter Extremism Project. (2020). *Extremist Content Online: ISIS Releases A New Video Titled "The Epic Battles Of Attrition 3" And Neo-Nazi Telegram Channel Removed.* Retrieved from https://www.counterextremism.com/press/extremist-content-online-isis-releases-new-video-titled-epic-battles-attrition-3-and-neo-nazi

Counter Extremism Project. (2020). *Roundup: Extremist Content Online*.

Counter Extremism Project. (2021). *Extremist Content Online: Online Extremists Create Digital Versions of Christchurch Terrorist Attack in Videogams*.

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

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

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

Curry, D. (2023, January). *Reddit Revenue and Usage Statistics (2023)*. Retrieved from https://www.businessofapps.com/data/reddit-statistics/

Dailymotion. (2023, June 1). *Prohibited Content Policy*. Retrieved from https://legal.dailymotion.com/en/terms-of-use/#prohibited-content

Dailymotion. (2023, February 28). *Transparency report on tackling terrorist and violent extremist content*.

Dang, S. (2023, July 31). *X reorganizes trust and safety team under Musk, CEO Yaccarino*. Retrieved from https://www.reuters.com/technology/x-reveals-reporting-structure-under-elon-musk-ceo-yaccarino-2023-07-31/

Datanyze. (2021). *File Sharing Software Market Share.* Retrieved from https://www.datanyze.com/market-share/file-sharing--198

Datareportal. (2023, July). *Global Social Media Statistics*. Retrieved from Datareportal: https://datareportal.com/social-media-users

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

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

Discord. (2021, May 7). *Announcing the Discord Moderator Academy Exam*. Retrieved from Discord Blog: https://blog.discord.com/announcing-the-discord-moderator-academy-exam-a1bcb5b9d405

Discord. (2021, October 7). *Discord Transparency Report: January - June 2021*. Retrieved from https://discord.com/blog/discord-transparency-report-h1-2021

Discord. (2021, May 25). *How Trust & Safety Addresses Violent Extremism on Discord*. Retrieved from Discord Blog: https://discord.com/blog/how-trust-safety-addresses-violent-extremism-on-discord

Discord. (2022). *Addressing harmful off platform behavior*. Retrieved from https://discord.com/safety/addressing-harmful-off-platform-behavior

Discord. (2022, June 16). *Meet your newest community moderator: AutoMod is here*.

Discord. (2022). *Our response to the tragedy in Buffalo*. Retrieved from https://discord.com/safety/our-response-to-the-tragedy-in-buffalo

Discord. (2023, March 27). *Community Guidelines*. Retrieved from https://discord.com/guidelines

Discord. (2023). *How we enforce rules - What actions we take*. Retrieved from https://discord.com/safety/360044159011-what-actions-we-take

Discord. (2023, March 30). *Transparency reports*. Retrieved from https://discord.com/safety-transparency-reports/2022-q4

Donovan, J., Lewis, B., & Friedberg, B. (2019). arallel Ports: Sociotechnical Change from the Alt-Right to Alt-Tech. In *In: Maik Fielitz, Nick Thurston (Hg.): Post-Digital Cultures of the Far Right: Online Actions and Offline Consequences in Europe and the US.* Bielefeld.

Douyin. (2023, September 8). *Douyin User Service Agreement*. Retrieved from https://www.douyin.com/draft/douyin_agreement/douyin_agreement_user.html?id=6773906068725565448

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

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

Element. (2023, November 10). *Element's legal - Terms of use*. Retrieved from https://element.io/legal

Envisage Digital. (2021). *WordPress Market Share in 2021*. Retrieved from Envisage Digital: https://www.envisagedigital.co.uk/wordpress-market-share/

eSafety Commissioner. (2021, September). *Industry codes and standards*. Retrieved from https://www.esafety.gov.au/industry/codes#esafety-position-paper

European Commission. (2023). *How the Digital Services Act enhances transparency online*. Retrieved from https://digital-strategy.ec.europa.eu/en/policies/dsa-brings-transparency#:~:text=to%20the%20users.-,Transparency%20reports,reports%20on%20their%20content%20moderation.

European Commission. (2023, January). *Terrorist content online: Commission takes action to protect people from the risk of online radicalisation and recruitment by extremists*. Retrieved from https://ec.europa.eu/commission/presscorner/detail/en/ip_23_132

European Parliament; European Council. (2017). *Directive (EU) 2017/541 of the European Parliament and of the Council of 15 March 2017 on combating terrorism and replacing Council Framework Decision 2002/475/JHA and amending Council Decision 2005/671/JHA*. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32017L0541

European Union. (2021). *Regulation (EU) 2021/784 of the European Parliamant and of the Council of 29 April 2021 on addresing the dissemination of terrorist content online*. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX%3A32021R0784

Europol . (2021). *Jihadist content targetd on Internet Archive platform*.

Europol. (2016, January 18). *Changes in modus operandi of Islamic State terrorist attacks*. Retrieved from https://www.europol.europa.eu/sites/default/files/documents/changes_in_modus_operandi_of_is_in_terrorist_attacks.pdf

EUROPOL. (2018, March). *More than 900 instances of online terrorist propaganda uncovered*. Retrieved from https://www.europol.europa.eu/media-press/newsroom/news/more-900-instances-of-online-terrorist-propaganda-uncovered

EUROPOL. (2021). *Terrorists attempted to take advantage of the pandemic, says Europol's new EU Terrorism Situation and Trend Report 2021*. Retrieved from https://www.europol.europa.eu/media-press/newsroom/news/terrorists-attempted-to-take-advantage-of-pandemic-says-europol%E2%80%99s-new-eu-terrorism-situation-and-trend-report-2021

EUROPOL. (2022, December). *14 countries tackle violent extremism online in a coordinated referral action day*. Retrieved from https://www.europol.europa.eu/media-press/newsroom/news/14-countries-tackle-violent-extremism-online-in-coordinated-referral-action-day

EUROPOL. (2022). *European Union Terrorism Situation and Trend Report* . Retrieved from https://www.europol.europa.eu/cms/sites/default/files/documents/Tesat_Report_2022_0.pdf

EUROPOL. (2023). *Terrorism situation and trend report*. Retrieved from https://www.europol.europa.eu/cms/sites/default/files/documents/European%20Union%20Terrorism%20Situation%20and%20Trend%20report%202023.pdf

Facebook. (2021). *Violence and Incitement*. Retrieved from https://transparency.fb.com/en-gb/policies/community-standards/violence-incitement/

Facebook. (2022, November 18). *Appealed content*. Retrieved from https://transparency.fb.com/policies/improving/appealed-content-metric/

Facebook. (2022, October 4). *Content actioned*. Retrieved from https://transparency.fb.com/policies/improving/content-actioned-metric/

Facebook. (2022, October 4). *Content actioned*. Retrieved from Facebook : https://transparency.fb.com/policies/improving/content-actioned-metric/

Facebook. (2022, January 19). *How review teams work*. Retrieved from Facebook: https://transparency.fb.com/en-gb/enforcement/detecting-violations/how-review-teams-work/

Facebook. (2022, January 26). *How Technology Detects Violations*. Retrieved from Facebook: https://transparency.fb.com/enforcement/detecting-violations/technology-detects-violations/

Facebook. (2022, November 18). *Prevalence*. Retrieved from Facebook: https://transparency.fb.com/en-gb/policies/improving/prevalence-metric/

Facebook. (2022, October). *Removing pages and groups*. Retrieved from https://transparency.fb.com/enforcement/taking-action/removing-pages-groups

Facebook. (2022, October 4). *Restored content*. Retrieved from https://transparency.fb.com/policies/improving/restored-content-metric/

Facebook. (2023). *Community Standards Enforcement Report.* Retrieved from Facebook: https://transparency.fb.com/data/community-standards-enforcement/?from=https%3A%2F%2Ftransparency.facebook.com%2Fcommunity-standards-enforcement%2Fguide

Facebook. (2023, July). *Community Standards, Dangerous Individuals and Organisations.* Retrieved from Facebook Transparency Center: https://transparency.fb.com/en-gb/policies/community-standards/dangerous-individuals-organizations/

Facebook. (2023). *Difference between an admin and moderator of a Facebook group*. Retrieved from https://www.facebook.com/help/901690736606156

Facebook. (2023, January). *Disabling accounts*. Retrieved from Facebook: https://transparency.fb.com/enforcement/taking-action/disabling-accounts

Facebook. (2023, July). *Facebook Community Standards*. Retrieved from https://transparency.fb.com/en-gb/policies/community-standards/violent-graphic-content/

Facebook. (2023). *Facebook Community Standards*. Retrieved from https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/

Facebook. (2023). *Help Centre - I don't think Facebook should have taken down my post.* Retrieved from https://www.facebook.com/help/2090856331203011?helpref=related_articles

Facebook. (2023). *Help Centre - What is the Oversight Board?* Retrieved from https://www.facebook.com/help/711867306096893?helpref=faq_content

Facebook. (2023, February 22). *Proactive rate*. Retrieved from https://transparency.fb.com/policies/improving/proactive-rate-metric/

Facebook. (2023, February). *Regulatory and Other Transparency Reports*. Retrieved from https://transparency.fb.com/data/regulatory-transparency-reports/

Facebook. (2023, February). *Restricting Accounts*. Retrieved from https://transparency.fb.com/en-gb/enforcement/taking-action/counting-strikes/

Facebook. (2023). *Taking down violating content*. Retrieved from https://transparency.fb.com/en-gb/enforcement/taking-action/taking-down-violating-content/

Feldstein, S., & Gordon, S. (2021). *Are Telegram and Signal Havens for Right-Wing Extremists?* .

Fernandez-Aubert, E.-J., Reinhart, R., & Squire, M. (2023, November 6). *Digital threat report: Odysee*. Retrieved from https://www.splcenter.org/hatewatch/2023/11/06/digital-threat-report-odysee

Fisher, A., Prucha, N., & Winterbotham, E. (2019). *Mapping the Jihadist Information Ecosystem: Towards the Next Generation of Disruption Capability .*

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

Fishwick, C. (2014). *How a Polish student's website became an Isis propaganda tool*. Retrieved from https://www.theguardian.com/world/2014/aug/15/-sp-polish-man-website-isis-propaganda-tool

Frenkel, S. (2022, October 28). *Elon Musk moves to form a content moderation council for Twitter.* Retrieved from https://www.nytimes.com/2022/10/28/technology/twitter-elon-musk-content-moderation.html

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

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

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

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

Gab. (2023). *Gab's Terms of Service*. Retrieved from https://gab.com/about/tos

Galov, N. (2023, May). *9 Post-Worthy QZone Statistics*. Retrieved from https://webtribunal.net/blog/qzone-statistics/#gref

GIFCT. (2022). *GIFCT Transparency Report*. Retrieved from https://gifct.org/wp-content/uploads/2022/12/GIFCT-Transparency-Report-2022.pdf

GIFCT. (2023, June 29). Retrieved from https://gifct.org/2023/06/29/borderline-content-understanding-the-gray-zone/

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

Google. (2023). *How Google autocomplete predictions work*. Retrieved from https://support.google.com/websearch/answer/7368877?sjid=15998257637882839334-EU#zippy=%2Cautocomplete-policies

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

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

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

Google. (n.d.). *Report a violation* . Retrieved from Google: https://support.google.com/docs/answer/2463296?hl=en&ref_topic=1360897&sjid=5436358909425045284-EU#zippy=%2Cgoogle-drive

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

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

Google/YouTube. (2023). *Appeal Community Guidelines actions.* Retrieved from Google, Youtube: https://support.google.com/youtube/answer/185111?hl=en

Google/YouTube. (2023). *Community Guidelines - How does YouTube identify content that violates the Community Guidelines?* Retrieved from https://www.youtube.com/intl/ALL_in/howyoutubeworks/policies/community-guidelines/#detecting-violations

Google/YouTube. (2023). *Community Guidelines strike basics - YouTube Help*. Retrieved from Google/YouTube: https://support.google.com/youtube/answer/2802032

Google/YouTube. (2023). *Disable or enable Restricted/Safe Mode.* Retrieved from Google, Youtube: https://support.google.com/youtube/answer/174084?hl=en&sjid=17414125458760940771-EU

Google/YouTube. (2023). *Limited features for certain videos - YouTube Help*. Retrieved from Google/YouTube: https://support.google.com/youtube/answer/7458465

Google/YouTube. (2023). *YouTube Community Guidelines enforcement - Violent Extremism*. Retrieved from Google/YouTube: https://transparencyreport.google.com/youtube-policy/featured-policies/violent-extremism?hl=en_GB&policy_removals=period:Y2019Q2&lu=policy_removals

Google/YouTube. (2023). *YouTube Community Guidelines Enforcement FAQs*. Retrieved from https://support.google.com/transparencyreport/answer/9209072#zippy=%2Chow-is-violative-view-rate-vvr-calculated

Google/YouTube. (2023). *YouTube Trusted Flagger program.* Retrieved from Google, Youtube: https://support.google.com/youtube/answer/7554338?&ref_topic=2803138

Gorwa, R. (2019). The platform governance triangle: conceptualising the informal regulation of online content. *Internet Policy Review, 8*(2).

Goujard, C. (2022, June 7). *Online platforms now have an hour to remove terrorist content in the EU*. Retrieved from https://www.politico.eu/article/online-platforms-to-take-down-terrorist-content-under-an-hour-in-the-eu/

Government of Austria. (2021). *Federal Act on measures to protect users on communication platforms (Communication Platforms Act)*. Retrieved from https://www.ris.bka.gv.at/Dokumente/Erv/ERV_2020_1_151/ERV_2020_1_151.html

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

Government of New Zealand. (2021). *Films, Videos, and Publications Classification (Urgent Interim Classification of Publications and Prevention of Online Harm) Amendment Bill*. Retrieved from https://www.legislation.govt.nz/bill/government/2020/0268/latest/LMS294551.html

Government of the United Kingdom. (2023). *Online Safety Act 2023*. Retrieved from https://bills.parliament.uk/bills/3137

Graham, R. (2016). How Terrorists Use Encryption . *CTCSENTINEL, 9*(6). Retrieved from https://ctc.westpoint.edu/how-terrorists-use-encryption/

Grayson, N. (2022). *How Twitch took down Buffalo shooter's stream in under two minutes*. Retrieved from The Washington Post: https://www.washingtonpost.com/video-games/2022/05/20/twitch-buffalo-shooter-facebook-nypd-interview/

Hale, J. (2022, April 29). *YouTube is axing all individual volunteers from its Trusted Flagger program*. Retrieved from Tubefilter: https://www.tubefilter.com/2022/04/29/youtube-trusted-flagger-program-individuals-organizations/

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

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

Hemjrajani, A. (2022, October 19). *CO22102 | The Indian Government Ban on Chinese Apps and the Singapore Connection*. Retrieved from https://www.rsis.edu.sg/rsis-publication/cens/the-indian-government-ban-on-chinese-apps-and-the-singapore-connection/

Hill QC, M. (February 2018). The Westminster bridge terrorist attack - A report on the use of terrorism legislation. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/695304/IRTL-Westminster_Bridge_Attack_Report_March_2018..pdf

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

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

Hymas, C. (2019, May 11). *Isil extremists using Instagram to promote jihad and incite support for terror attacks on the West* . Retrieved from The Telegraph:

https://www.telegraph.co.uk/news/2019/05/11/isil-extremists-using-instagram-promote-jihad-incite-support/

Ilinsky, A. (2019, October 21). *Interview with Mariusz Zurawek, founder of JustPaste.it, the anonymous sharing tool*. Retrieved from https://hostadvice.com/blog/justpaste-it-is-the-quickest-way-to-share-content-online/

IMO. (2023, May 17). *Community Guidelines*. Retrieved from https://imo.im/policies/community_guidelines.html

Instagram. (2023, February 28). *European Union Terrorist Content Online Transparency Report*. Retrieved from https://transparency.fb.com/sr/eu-online-report-ig-feb28-23

Instagram. (2023). *Help Center*.

Instagram. (2023). *Terms and Policies- Community Guidelines*.

Instagram. (2023). *Terms and Policies- Dissemination of Terrorist Content Online*.

Instagram. (2023). *Threads Terms of Use*. Retrieved from https://help.instagram.com/769983657850450/?helpref=uf_share

Instagram. (2023). *Threads Terms of Use*.

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

Josh. (2022, November 7). *Transparency Report*. Retrieved from https://share.myjosh.in/transparency-report?lang=kn

Josh. (2023). *Disclosure of Grievance Details by the Intermediary*. Retrieved from https://share.myjosh.in/grievance-data

Josh. (2023, February). *JOSH Terms of Service*. Retrieved from https://share.myjosh.in/terms-conditions?lang=en

Justepaste.it. (2018, March 20). *Anonymous by default*. Retrieved from https://justpaste.it/1ikbg

Justpaste.it. (2020). *JustPaste.it Transparency Report 2020.* Retrieved from https://justpaste.it/transparency_report_2019

Justpaste.it. (2021). *JustPaste.it Transparency Report 2020.* Retrieved from https://justpaste.it/transparency_report_2020

Justpaste.it. (2022). *JustPaste.it Transparency Report 2022*. Retrieved from https://justpaste.it/transparency_report_2022

JustPaste.it. (2023). *Content/ account blockage appeal*. Retrieved from https://justpaste.it/terms/appeal

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

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

Keller, D., & Leerssen, P. (2020). Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation. In *in Nathaniel Persily and Joshua A. Tucker (eds) Social Media and Democracy, The State of the Field, Prospects for Reform.* Cambridge University Press.

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

King, J. (2020, August 11). *How We Review Content*. Retrieved from Facebook : https://about.fb.com/news/2020/08/how-we-review-content/

Knight, W. (2017, January 26). *The Insanely Popular Chinese News App That You've Never Heard Of*. Retrieved from https://www.technologyreview.com/2017/01/26/154363/the-insanely-popular-chinese-news-app-that-youve-never-heard-of/

Knockel, J., & Ruan, L. (2021, August). *Measuring QQMail's automated email censorship in China*. Retrieved from https://dl.acm.org/doi/10.1145/3473604.3474560

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

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

Kohlmann, E. F. (2015). *Charlie Hebdo and the Jihadi Online Network: Assessing the Role of American Commercial Social Media Platforms.* Retrieved from https://docs.house.gov/meetings/FA/FA18/20150127/102855/HHRG-114-FA18-Wstate-KohlmannE-20150127.pdf

Kuaishou. (2023, May). *Kuaishou Technology Announces First Quarter 2023 Unaudited Financial Results*. Retrieved from https://ir.kuaishou.com/news-releases/news-release-details/kuaishou-technology-announces-first-quarter-2023-unaudited/#:~:text=Average%20MAUs%20on%20Kuaishou%20APP,the%20same%20period%20of%202022.

Kuaishou. (2023). *Live permission*. Retrieved from https://www.kuaishou.com/help/feedback/2706?categoryId=2670&subCategoryId=2675

Kuhaishou/Kwai. (2022). *Transparency Report 1H 2022*. Retrieved from https://www.kwai.com/safety/resources?id=transparency

Kulvi, F. (2021, July 8). *Meet Rumble, Canada's new 'free speech' platform — and its impact on the fight against online misinformation*. Retrieved from The Conversation: https://theconversation.com/meet-rumble-canadas-new-free-speech-platform-and-its-impact-on-the-fight-against-online-misinformation-163343

Kwai. (2022, February). *Community Guidelines*. Retrieved from https://www.kwai.com/safety?id=community

Kwai. (2022, January). *Terms of Service*. Retrieved from https://app.kwai.com/agreement/service-terms

Lakhani, S. (2021). *Video Gaming and (Violent) Extremism - An exploration of the current landscape, trends, and threats*. Retrieved from https://home-affairs.ec.europa.eu/system/files/2022-02/EUIF%20Technical%20Meeting%20on%20Video%20Gaming%20October%202021%20RAN%20Policy%20Support%20paper_en.pdf

Lakhani, S. (2021). *Video Gaming and (Violent) Extremism: An exploration of the current landscape, trends, and threats*. Retrieved from https://home-affairs.ec.europa.eu/system/files/2022-02/EUIF%20Technical%20Meeting%20on%20Video%20Gaming%20October%202021%20RAN%20Policy%20Support%20paper_en.pdf

Lamphere-Englund, G., & White, D. J. (2022, May 16). *The Buffalo Attack and the Gamification of Violence*. Retrieved from https://www.rusi.org/explore-our-research/publications/commentary/buffalo-attack-and-gamification-violence

Lamphere-Englund, G., & White, J. (2023, May). *The Online Gaming Ecosystem:: Assessing Digital Socialisation, Extremism Risks and Harms Mitigation Efforts*. Retrieved from https://gnet-research.org/wp-content/uploads/2023/05/GNET-37-Extremism-and-Gaming_web.pdf

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

Lawit, B., & Xu, Y. (2023, February 22). *Sharing LinkedIn's Responsible AI Principles*. Retrieved from https://blog.linkedin.com/2023/february/22/responsible-ai-principles

Le Journal du Dimanche. (2015, July 19). *Terrorisme : le casse-tête des messageries cryptées*. Retrieved from https://www.lejdd.fr/societe/terrorisme-le-casse-tete-des-messageries-cryptees-29541

Leidig, D. E. (2021, February 17). *Odysee: The New YouTube for the Far-Right*. Retrieved from https://gnet-research.org/2021/02/17/odysee-the-new-youtube-for-the-far-right/

Liao, R. (2019, March 21). *PicsArt hits 130 million MAUs as Chinese flock to its photo-editing app*. Retrieved from Tech Crunch: https://techcrunch.com/2019/03/20/picsart-china/

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

LINE. (2020, September 9). *LINE Becomes 1st Asia-Based Company to Join Christchurch Call*. Retrieved from https://linecorp.com/en/pr/news/en/2020/3381

LINE. (2022). *LINE Content Moderation Report*. Retrieved from https://linecorp.com/en/security/moderation/2022h1

LINE. (2023). *Transparency Report*. Retrieved from https://linecorp.com/en/security/transparency/top

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

LinkedIn. (2022). *Dangerous Organisations and Individuals*. Retrieved from https://www.linkedin.com/help/linkedin/answer/a1342780?lang=en-US

LinkedIn. (2023). *Create and host LinkedIn Live: Access criteria*. Retrieved from https://www.linkedin.com/help/linkedin/answer/a568503

LinkedIn. (2023). *How we enforce our Professional Community Policies*. Retrieved from https://www.linkedin.com/help/linkedin/answer/a1342754

LinkedIn. (2023). *Types of content policy violations*. Retrieved from https://about.linkedin.com/transparency/community-report/content-violations

Lix Xan Wong, K., & Shields Dobson, A. (2019). We're just data: Exploring China's social credit system in relation to digital platform ratings cultures in Westernised democracies. *Global Media and China, 4*(2), 220-232.

Lu, S. (2021, February 18). *I helped build ByteDance's censorship machine*. Retrieved from https://www.protocol.com/china/i-built-bytedance-censorship-machine

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

Lyons, K. (2022, November). *28 Top Social Media Platforms Worldwide*. Retrieved from https://www.semrush.com/blog/most-popular-social-media-platforms/

Mandav, J., Parihar, R., Saket, S., Gupta, V., & Mukherjee, D. (2021, May 26). *Multimodal Automated Content Moderation*. Retrieved from https://medium.com/sharechat-techbyte/multimodal-automated-content-moderation-69876e6a9d85

Marketing to China. (2021, July). *Guide to Douban Marketing*. Retrieved from https://marketingtochina.com/guide-to-douban-marketing/#:~:text=With%20over%20300%20million%20monthly,using%20the%20app%20without%20registering.

Maslar, H. (2022, March). *Viber's formula to measure active users*. Retrieved from https://mixpanel.com/blog/metrics-that-matter-to-viber-a-formula-to-meaningfully-measure-active-users/

Mastodon. (2023, October 2). *Annual Report 2022*. Retrieved from https://blog.joinmastodon.org/2023/10/annual-report-2022/

Mastodon. (2023, March 17). *Dealing with unwanted content*. Retrieved from https://docs.joinmastodon.org/user/moderating/#report

Mastodon. (n.d.). *About mastodon.social* . Retrieved from https://mastodon.social/about

Mastodon.social. (2023, December 7). *Moderation actions*. Retrieved from https://docs.joinmastodon.org/admin/moderation/

Matrix. (2023). *Code of Conduct*. Retrieved from https://matrix.org/legal/code-of-conduct/

Matrix. (2023, January 25). *Community moderation*. Retrieved from https://matrix.org/docs/communities/moderation/

McKinsey & Company. (2020). *How COVID-19 has pushed companies over the technology tipping point—and transformed business forever*. Retrieved from https://www.mckinsey.com/capabilities/strategy-and-corporate-finance/our-insights/how-covid-19-has-pushed-companies-over-the-technology-tipping-point-and-transformed-business-forever

Mega.nz. (2021, September). *Mega Transparency Report - September 2021*. Retrieved from https://mega.io/Mega_Transparency_Report_September_2021.pdf

Mega.nz. (2021, October 20). *Mega Transparency Report 2021 (blogpost)*. Retrieved from https://mega.io/blog/mega-transparency-report-2021

Meihan, L. (2022, February 25). *China's Content Moderators Are Overworked and Chronically Stressed*. Retrieved from https://www.sixthtone.com/news/1009742

Meihan, L. (2022, February 25). *China's content moderators are overworked and chronically stressed*. Retrieved from https://www.sixthtone.com/news/1009742

MEMRI. (2023). *ISIS Supporters Disseminates French-Language List of ISIS Materials Hosted on San-Francisco Based Internet Archive*.

Meta. (2023, May 12). *Transparency Center - Reviewing high-impact content accurately via our cross-check system*. Retrieved from https://transparency.fb.com/enforcement/detecting-violations/reviewing-high-visibility-content-accurately/

Meta. (2023). *Transparency Center- Complaint Handling Process*.

Meta. (2023). *Transparency Center- Facebook Community Standards*.

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

Microsoft. (2020-2022). *Digital Safety Content Report*. Retrieved from https://www.microsoft.com/en-us/corporate-responsibility/digital-safety-content-report?activetab=pivot_1:primaryr4

Microsoft. (2023, January). *Microsoft Fiscal Year 2023 Second Quarter Earnings Conference Call*. Retrieved from https://www.microsoft.com/en-us/Investor/events/FY-2023/earnings-fy-2023-q2.aspx

Microsoft. (2023). *Report Terrorist Content Posted to a Microsoft Consumer Service*. Retrieved from https://www.microsoft.com/en-us/concern/terroristcontent

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

Middle East Eye. (2015). *Islamic State supporters set up Facebook rival*. Retrieved from https://www.middleeasteye.net/news/islamic-state-supporters-set-facebook-rival

Middle East Media Research Institute. (2014). *Hosted In Germany, Justpaste.it Is Being Widely Used By Terrorist Organizations To Publish Jihadist Content*. Retrieved from https://www.memri.org/cjlab/hosted-in-germany-justpaste-it-is-being-widely-used-by-terrorist-organizations-to-publish-jihadist-content

Moj. (2022, July 28). *Content and Community Guidelines*. Retrieved from https://help.mojapp.in/policies/content-policy/

Moj. (2022, July 28). *Terms of Use*. Retrieved from https://help.mojapp.in/policies/terms

Moj. (2023). *Transparency Report*. Retrieved from https://help.mojapp.in/transparency-report/?q=moj-july-2023

Mostofa, S. M. (2022). *Cyber Radicalization by Bangladeshi Islamists* .

Naffakh, M. (2022, November 25). *How pro-terrorism accounts are circumventing moderation on social media*. Retrieved from https://observers.france24.com/en/middle-east/20221125-social-media-propaganda-islamic-state-terrorism

Naveen , P. (2023). *Hizb-ut-Tahrir suspects nabbed by Madhya Pradesh ATS used 'Threema', 'Rocket Chat' to communicate*. Retrieved from https://timesofindia.indiatimes.com/city/bhopal/hizb-ut-tahrir-suspects-nabbed-by-madhya-pradesh-ats-used-threema-rocket-chat-to-communicate/articleshow/100172315.cms

New America. (n.d.). *How Automated Tools are Used in the Content Moderation Process*. Retrieved from https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/how-automated-tools-are-used-in-the-content-moderation-process/

Newman, L. H. (2023, June 5). *Apple Expands Its On-Device Nudity Detection to Combat CSAM*. Retrieved from https://www.wired.com/story/apple-communication-safety-nude-detection/

Noack, R., Beck, L., & Morris, L. (2019). *Gunman live-streamed attack outside German synagogue that left two dead*. Retrieved from The Washington Post: https://www.washingtonpost.com/world/shooting-near-synagogue-in-germany-leaves-at-least-two-people-dead-police-say/2019/10/09/08214514-ea89-11e9-9306-47cb0324fd44_story.html

O'Connor, C., & Smith, M. (2023). *It is (still) shockingly easy to find terrorist content on TikTok*. Retrieved from https://www.isdglobal.org/digital_dispatches/it-is-still-shockingly-easy-to-find-terrorist-content-on-tiktok/

O'Connor, C., & Smith, M. (2023). *It is (still) shockingly easy to find terrorist content on TikTok*. Retrieved from https://www.isdglobal.org/digital_dispatches/it-is-still-shockingly-easy-to-find-terrorist-content-on-tiktok/

Odysee. (2021, March 3). *Moderation Tools*. Retrieved from https://odysee.com/@OdyseeHelp:b/moderation:f

Odysee. (2022, February 16). *Declaration of Indifference: Community Guidelines*. Retrieved from https://help.odysee.tv/communityguidelines/

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

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

OECD. (2022, November). *Putting people first in digital transformation*. Retrieved from https://www.oecd-ilibrary.org/docserver/865f8426-en.pdf?expires=1696861716&id=id&accname=ocid84004878&checksum=1E8293B99E478E31B2C87C9089B6FF5D

OECD. (2022). *Transparency reporting on terrorist and violent extremist content online 2022*. Retrieved from https://www.oecd.org/digital/transparency-reporting-on-terrorist-and-violent-extremist-content-online-2022-a1621fc3-en.htm#:~:text=This%20is%20the%20third%20benchmarking,2020%20and%20eleven%20in%202021.

Ofcom. (2022, October 12). *The Buffalo Attack: Implications for Online*.

Oversight Board. (2022). *Oversight Board overturns Meta's original decision in 'Mention of the Taliban in news reporting'* . Retrieved from https://www.oversightboard.com/news/484790580170915-oversight-board-overturns-meta-s-original-decision-in-mention-of-the-taliban-in-news-reporting-2022-005-fb-ua/

Perez, S. (2022, May 4). *Pinterest quietly launches a livestreaming app for video creators*. Retrieved from https://techcrunch.com/2022/05/04/pinterest-quietly-launches-a-live-streaming-app-for-video-creators/

Picsart. (2015). *Picsart Community Guidelines Blogpost*. Retrieved from Picsart Blog: https://picsart.com/blog/post/picsart-community-guidlines

Picsart. (2023). *Community Guidelines and Spaces Rules*. Retrieved from https://support.picsart.com/hc/en-us/articles/11000111767581-Community-Guidelines-and-Spaces-Rules

Picsart. (2023). *Space Owner & Admin Code of Conduct*. Retrieved from https://picsart.com/admin-code-of-conduct

Picsart. (n.d.). *How do I report inappropriate behavior or content on Picsart?* Retrieved from Picsart Help Center: https://support.picsart.com/hc/en-us/articles/360003824257-How-do-I-report-inappropriate-behavior-or-content-on-Picsart-

Pinterest. (2023). *Community Guidelines*. Retrieved from https://policy.pinterest.com/en-gb/community-guidelines

Pinterest. (2023). *Enforcement*. Retrieved from https://policy.pinterest.com/en/enforcement

Pinterest. (2023). *Livestream on Pinterest TV*. Retrieved from https://help.pinterest.com/en/business/article/livestream-on-pinterest-tv

Pinterest. (2023). *Transparency Report*. Retrieved from https://policy.pinterest.com/en/transparency-report

Pinto, N. T. (2018). *The Portugal Connection in the Strasbourg-Marseille Islamic State Terrorist Network.* Retrieved from https://ctc.westpoint.edu/portugal-connection-strasbourg-marseille-islamic-state-terrorist-network/

Provetti, A. (2021). *Live Monitoring 4chan Discussion Threads*. Retrieved from https://air.unimi.it/handle/2434/890320

Quora. (2021). *How do I appeal a Quora Moderation decision?* Retrieved from https://help.quora.com/hc/en-us/articles/360000480343-How-do-I-appeal-a-Quora-Moderation-decision-

Quora. (n. d.). *Brand Safety on Quora*. Retrieved from https://go.quoraforbusiness.com/rs/384-CMP-465/images/Brand_Safety_on_Quora.pdf

r/pan. (2022, November). *Update on the future of live video broadcasting on Reddit*. Retrieved from https://www.reddit.com/r/pan/comments/yl5zzd/update_on_the_future_of_live_video_broadcasting/

Radsch, C. (2023, April). *Transparency Reporting: Good Practices and Lessons from Global Assessment Frameworks*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4416400

Ragot, J. (2023, May). *BFM TV - La player Dailymotion est toujours utilisée par 90% des internautes français (sans pour autant qu'ils le sachent)*. Retrieved from https://www.bfmtv.com/tech/actualites/streaming/dailymotion-est-toujours-utilise-par-9-francais-sur-10_AV-202305160266.html

Ray, S. (2021). *The Far-Right Is Flocking To These Alternate Social Media Apps — Not All Of Them Are Thrilled*. Retrieved from https://www.forbes.com/sites/siladityaray/2021/01/14/the-far-right-is-flocking-to-these-alternate-social-media-apps---not-all-of-them-are-thrilled/?sh=1f7fa00f55a4

Reddit. (2021). *Quarantined Subreddits*. Retrieved from Reddit Help: https://www.reddithelp.com/en/categories/rules-reporting/account-and-community-restrictions/quarantined-subreddits

Reddit. (2023). *2023 Transparency Report*. Retrieved from https://www.redditinc.com/policies/2023-h1-transparency-report.

Reddit. (2023). *Content Policy*. Retrieved from https://support.reddithelp.com/hc/en-us/articles/19003525756564

Reddit. (2023). *How does Reddit fight the dissemination of terrorist content?* Retrieved from https://support.reddithelp.com/hc/en-us/articles/19003525756564-How-does-Reddit-fight-the-dissemination-of-terrorist-content

Reddit. (2023). *Mod Help Centre*. Retrieved from https://mods.reddithelp.com/hc/en-us

Reddit. (2023). *Reddit Transparency Report 1H 2023*. Retrieved from https://www.redditinc.com/policies/2023-h1-transparency-report

Reddit Inc. (2022, September 8). *Moderator Code of Conduct*. Retrieved from Reddit: https://www.redditinc.com/policies/moderator-code-of-conduct

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

Reuters. (2022, April 28). *China's Weibo shows user locations to combat 'bad behaviour'*. Retrieved from Reuters: https://www.reuters.com/world/china/weibo-shows-user-locations-combat-bad-behaviour-2022-04-28/

Robertson , A. (2020). *FBI arrests alleged member of prolific neo-Nazi swatting ring* .

Rocket.Chat. (2019, September 6). *Code of Conduct: Services*. Retrieved from https://docs.rocket.chat/customer-center/legal-center/code-of-conduct-services

Rocket.Chat. (2020, September 24). *Censorship and Harmful Content*. Retrieved from https://docs.rocket.chat/customer-center/legal-center/law-enforcement/censorship-and-harmful-content

Rocket.Chat. (2023, November 30). *Acceptable Use Policy*. Retrieved from https://docs.rocket.chat/applicable-terms/supplemental-terms/acceptable-use-policy

Rocket.Chat. (2023, November 30). *Terms of Use*. Retrieved from https://docs.rocket.chat/applicable-terms/supplemental-terms/terms-of-use

Roy, S., & Mishra, D. (2023, February). *Interview | Focus is on profitability, growth at all costs never sustainable: ShareChat CEO*. Retrieved from https://economictimes.indiatimes.com/tech/startups/exclusive-focus-is-on-profitability-growth-at-all-costs-never-sustainable-sharechat-ceo/articleshow/98130762.cms?from=mdr

Ruan, L. (2019, October 7). *Regulation of the internet in China: An explainer*. Retrieved from https://theasiadialogue.com/2019/10/07/regulation-of-the-internet-in-china-an-explainer/

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

Rumble. (2023). *Terms and Conditions of Use and Agency Agreement*. Retrieved from https://rumble.com/s/terms

Rumble. (2024). *Our Story*. Retrieved from https://corp.rumble.com/our-story/

Santa Clara University's High Tech Law Institute. (2021). *The Santa Clara Principles On Transparency and Accountability in Content Moderation 2.0*. Retrieved from santaclaraprinciples.org: https://santaclaraprinciples.org/

Sarang, V. (2022, May 17). *Community Standards Enforcement Report Assessment Results*. Retrieved from https://about.fb.com/news/2022/05/community-standards-enforcement-report-assessment-results/

Scott, M. (2021, October 25). *Facebook did little to moderate posts in the world's most violent countries*. Retrieved from https://www.politico.eu/article/facebook-content-moderation-posts-wars-afghanistan-middle-east-arabic/

Scott, M. (2022, February). *Islamic State evolves 'emoji' tactics to peddle propaganda online*. Retrieved from https://www.politico.eu/article/islamic-state-disinformation-social-media/

Sebbagh, D. (2021, October). Facebook trained its AI to block violent live streams after Christchurch attacks. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2021/oct/29/facebook-trained-its-ai-to-block-violent-live-streams-after-christchurch-attacks

ShareChat. (2021, June 16). *Content Policy*. Retrieved from https://help.sharechat.com/policies/content-policy/

ShareChat. (2021, March 10). *Groups Policy*. Retrieved from https://help.sharechat.com/policies/groups-policy

ShareChat. (2021, December 17). *Terms of Use*. Retrieved from https://help.sharechat.com/policies/terms/

ShareChat. (2023). *Transparency Report (1st - 31 July 2023)*. Retrieved from https://help.sharechat.com/transparency-report/?q=sharechat-july-2023

SignHouse. (2023). *iMessage Revenue and Growth Statistics (2023)*. Retrieved from https://www.usesignhouse.com/blog/imessage-stats#:~:text=iMessage%20Monthly%20Active%20Users,-Want%20a%20link&text=iMessage%20has%20approximately%201%20billion%20monthly%20active%20users.,re%20sending%20messages%20or%20not.

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

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

Slack. (2021). *Acceptable Use Policy*. Retrieved from https://slack.com/acceptable-use-policy

Slack. (2021, September 1). *Community Forum Terms of Service* . Retrieved from https://slack.com/terms-of-service/community

Slack. (2024). *Workspace administration*. Retrieved from https://slack.com/help/articles/25273135315347-Review-flagged-content-on-Enterprise-Grid

Snapchat. (2021). *Transparency Report (July - December 2021)*. Retrieved from https://www.snap.com/en-US/privacy/transparency/2020-12-31?lang=en-US

Snapchat. (2022). *About Transparency Reporting*. Retrieved from https://values.snap.com/privacy/transparency/about

Snapchat. (2022, October 11). *Meet Snap's new Safety Advisory Board!* Retrieved from https://values.snap.com/news/meet-snaps-new-safety-advisory-board

Snapchat. (2023, January). *Community Guidelines*. Retrieved from https://values.snap.com/privacy/transparency/community-guidelines

Snapchat. (2023, March 15). *Content Guidelines for Recommendation Eligibility*. Retrieved from https://snap.com/en-GB/content-recommendation-guidelines#introduction

Snapchat. (2023, January). *Explainer - Hateful Content, Terrorism, and Violent Extremism*. Retrieved from https://values.snap.com/privacy/transparency/community-guidelines/hateful-content

Snapchat. (2023, August). *Snapchat Moderation, Enforcement, and Appeals*. Retrieved from https://values.snap.com/privacy/transparency/community-guidelines/moderation

Snapchat. (2023). *Transparency Report*. Retrieved from https://snap.com/en-GB/privacy/transparency

Snapchat. (2023). *Transparency Report Glossary*. Retrieved from https://values.snap.com/privacy/transparency/glossary

SoundCloud. (2023, December 18). *Community Guidelines*. Retrieved from https://soundcloud.com/community-guidelines

SoundCloud. (2024, January). *Content Moderation and Enforcement Policy*. Retrieved from https://soundcloud.com/moderation-and-enforcement-policies

Stackpole, T. (2022, November 9). *Content Moderation Is Terrible by Design*. Retrieved from https://hbr.org/2022/11/content-moderation-is-terrible-by-design

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

Statista. (2022). *Average number of monthly active users of Bilibili Inc. from 1st quarter 2019 to 1st quarter 2023*. Retrieved from https://www.statista.com/statistics/1109108/bilibili-average-monthly-active-users/

Statista. (2022, June). *LINE - statistics and facts*. Retrieved from https://www.statista.com/topics/1999/line/#topicOverview

Statista. (2022, March). *Number of monthly active users (MAU) of SVOD app Youku Tudou in China from June 2021 to March 2022*. Retrieved from https://www.statista.com/statistics/1277948/china-online-video-platform-youku-tudou-mobile-app-monthly-active-user-number/#:~:text=In%20March%202022%2C%20Youku%20Tudou's,active%20users%20to%20247%20million.

Statista. (2023, February). *Number of monthly active users of Xigua Video in China in selected months from November 2021 to December 2022*. Retrieved from https://www.statista.com/statistics/1364841/china-monthly-active-users-xigua-video/#:~:text=Xigua%20Video%2C%20one%20of%20the,the%20country%20in%20December%202022.

Statista. (2023). *Number of subscribing members of the Chinese online video site iQIYI from December 2017 to December 2022*. Retrieved from https://www.statista.com/statistics/1106180/china-online-video-platform-iqiyi-subscription-number/#:~:text=By%20the%20end%20of%202022,530%20million%20monthly%20active%20users.

Statista. (2023). *Worldwide visits to Zoom.us from November 2022 to April 2023*. Retrieved from https://www.statista.com/statistics/1259905/zoom-website-traffic/

Steam. (2023). *Reporting content within the Steam Community*. Retrieved from https://help.steampowered.com/en/faqs/view/4DE7-17AA-0E8B-C1AD

Steam. (n.d.). *Community Moderation*. Retrieved from https://partner.steamgames.com/doc/marketing/community_moderation?l=english

Steam. (n.d.). *Rules and Guidelines For Steam: Discussions, Reviews, and User Generated Content*. Retrieved from https://help.steampowered.com/en/faqs/view/6862-8119-C23E-EA7B

Stern, R. C. (2023). *LinkedIn Stats Looking Into 2023*. Retrieved from https://www.linkedin.com/pulse/linkedin-stats-looking-2023-robert-c-stern/

Stocking, G., Mitchell, A., Matsa, K. E., Widjaya, R., Jurkowitz, M., Ghosh, S., . . . Aubin, C. S. (2022, October). *Alternative social media sites frequently identify as free speech advocates*. Retrieved from https://www.pewresearch.org/journalism/2022/10/06/alternative-social-media-sites-frequently-identify-as-free-speech-advocates/

TamTam.Chat. (2022, April 15). *TamTam Chats and Channels Regulations*. Retrieved from https://about.tamtam.chat/en/regulations/

Tardi, C. (2022, July 20). *Monthly Active Users (MAU): Definition and How the Indicator Is Used*. Retrieved from https://www.investopedia.com/terms/m/monthly-active-user-mau.asp

tech against terrorism . (2021). *Terrorist Use of E2EE: State of play, misconceptions, and mitigation strategies REPORT.*

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

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

Tech Against Terrorism. (2021). *Knowledge Sharing Platform*. Retrieved from https://techagainstterrorism.org/knowledge-sharing-platform

Tech Against Terrorism. (2021, March). *Tech Against Terrorism Membership*. Retrieved from https://techagainstterrorism.org/news/2021/03/31/announcing-tech-against-terrorisms-newest-members-2

Tech Against Terrorism. (2021, July). *Trends in Terrorist and Violent Extremist Use of the Internet (Q1 – Q2 2021)*. Retrieved from https://www.techagainstterrorism.org/research

Tech Against Terrorism. (2021, July). *Trends in Terrorist and Violent Extremist Use of the Internet | Q1-Q2 2021*. Retrieved from https://www.techagainstterrorism.org/wp-content/uploads/2021/07/Tech-Against-Terrorism-Q1-Q2-TVEC-Trends-2021.pdf

Tech Against Terrorism. (2023). *Terrorist Use of Generative AI*. Retrieved from https://techagainstterrorism.org/gen-ai

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

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

The Economist. (2022, August 30). *The secrets of big tech*. Retrieved from https://www.economist.com/business/the-finance-secrets-of-big-tech/21808956

The Economist. (2023, August 2). *Can big tech keep getting bigger in the age of AI?* Retrieved from https://www.economist.com/leaders/2023/08/02/can-big-tech-keep-getting-bigger-in-the-age-of-ai

The London Economic. (2018). *Two men & teenager in court accused of being members of National Action in rebranded group called "Triple K Mafia"*.

The Straits Times. (2021, January 8). *Chinese video-sharing platform Douyin fined for pornographic, vulgar content*. Retrieved from https://www.straitstimes.com/asia/east-asia/chinese-video-sharing-platform-douyin-fined-for-pornographic-vulgar-content

Theema. (2022). *Terms and Conditions of Use of the Theema Server Service*. Retrieved from https://threema.ch/tos/?lang=en

TheWire. (2022). *JeM Financier Uses Social Media to Amplify Terrorist Propaganda Calling on Indian Muslims to Revolt*. Retrieved from https://thewire.in/tech/jaish-e-mohammed-terrorist-recruiter-propaganda-calls-on-indian-muslims-to-revolt

Threema. (2023). *Transparency Report*. Retrieved from https://threema.ch/en/transparencyreport

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

TikTok. (2023). *Community Guidelines Enforcement Report*. Retrieved from https://www.tiktok.com/transparency/en/community-guidelines-enforcement-2023-1/

TikTok. (2023). *Content violations and bans*. Retrieved from https://support.tiktok.com/en/safety-hc/account-and-user-safety/content-violations-and-bans

TikTok. (2023, March). *TikTok Community Guidelines - Safety and Civility*. Retrieved from https://www.tiktok.com/community-guidelines/en/safety-civility/

Tindall, R. (2022, September 24). *What are the current regulations for live streaming in China*. Retrieved from China-Britan Business Focus: https://focus.cbbc.org/what-are-the-current-regulations-for-live-streaming-in-china/

Tindall, R. (2022, September). *What are the current regulations for live streaming in China?* Retrieved from https://focus.cbbc.org/what-are-the-current-regulations-for-live-streaming-in-china/

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

Toutiao. (n.d.). *Toutiao User Agreement*. Retrieved from https://www.toutiao.com/user_agreement/

Trudeau, J. (2021, December 26). *Minister of Canadian Heritage Mandate Letter*. Retrieved from https://pm.gc.ca/en/mandate-letters/2021/12/16/minister-canadian-heritage-mandate-letter

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

Tumblr. (2013-2022). *Tumblr Transparency Report*. Retrieved from Tumblr: https://www.tumblr.com/transparency

Tumblr. (2022). *Community Guidelines*. Retrieved from https://www.tumblr.com/policy/en/community

Tumblr. (2023). *EU Terrorist Content Removal Orders*. Retrieved from https://transparency.automattic.com/tumblr/eu-terrorist-content-removal-orders/

Tumblr. (2023, August 3). *Introducing Tumblr Live*. Retrieved from https://staff.tumblr.com/post/724649961513533440/hello-tumblr-tumblr-here-were-launching-live

Tumblr. (2023, June 7). *Terms of Service*. Retrieved from https://www.tumblr.com/policy/fr/terms-of-service

Tumblr. (2023). *Tumblr Live Overview & FAQ*. Retrieved from https://help.tumblr.com/hc/en-us/articles/10073453888535-Tumblr-Live-Overview-FAQ

Tumblr Live. (2021, August 23). *Content and Conduct Policy*. Retrieved from https://terms.video.tumblr-live.com/content-and-conduct

Twitch. (2022). *H2 2022 Transparency Report*. Retrieved from Twitch: https://safety.twitch.tv/s/article/H2-2022-Transparency-Report?language=en_US#2H22022SafetyUpdates

Twitch. (2022). *Username Policy*. Retrieved from https://safety.twitch.tv/s/article/Usernames?language=en_US#:~:text=Usernames%20and%20display%20names%20created,Hateful%20Conduct

Twitch. (2023). *Community Guidelines*. Retrieved from https://www.twitch.tv/p/fr-fr/legal/community-guidelines/20210527/

Twitch. (2023). *Safety at Twitch*. Retrieved from https://safety.twitch.tv/s/article/Safety-at-Twitch?language=en_US

Twitter. (2012-2022). *X Transparency Rules enforcement*. Retrieved from Twitter Transparency Reportt: https://transparency.twitter.com/en/twitter-rules-enforcement.html

U.S. Department of Justice. (2020). *Global Disruption of Three Terror Finance Cyber-Enabled Campaigns*. Retrieved from https://www.justice.gov/usao-dc/pr/global-disruption-three-terror-finance-cyber-enabled-campaigns

UN CTED. (2021). Retrieved from https://www.un.org/securitycouncil/ctc/sites/www.un.org.securitycouncil.ctc/files/ctc_cted_factsheet_ct_in_cyberspace_oct_2021.pdf

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

US Treasury. (2023, August 3). *OFFICE OF FOREIGN ASSETS CONTROL - Specially Designated Nationals and Blocked Persons List.* Retrieved from Treasury: https://www.treasury.gov/ofac/downloads/sdnlist.pdf

Veilleux-Lepage, Y., Daymon, C., & Archambault, E. (2022). *Learning from Foes: How Racially and Ethnically Motivated Extremists Embrace and Mimic Islamic States' Use of Emerging Technologies.*

Verizon Media. (2019). *Transparency Report*. Retrieved from Verizon Media: https://www.verizonmedia.com/transparency/index.html?guce_referrer=aHR0cHM6Ly90cmFuc3BhcmVuY3kub2F0aC5jb20vaW5kZXguaHRtbD9ndWNlX3JlZmVycmVyPWFlUjBjSE02THk5M2QzY3VkSFZ0WW14eUxtTnZiUzgmZ3VjZV9yZWZlcnJlcl9zaWc9QVFBQUFKazduZ3VNWS04dHhtNG9hWFM3TUlkNkxlUWxkMEZ5

Verma, M. (2023). *India has blocked 14 mobile messenger apps on securiy fears*.

Viber. (2022, March 22). *Viber Acceptable Use Policy*. Retrieved from https://www.viber.com/en/terms/viber-public-content-policy/

Viber. (2023). *Communities Knowledge Base For Admins*. Retrieved from https://help.viber.com/hc/en-us/articles/9340728226077-Communities-Knowledge-Base-For-Admins

Viber. (2023, April 17). *Viber Terms of Service*. Retrieved from https://www.viber.com/en/terms/viber-terms-use/

Vimeo. (2023). *How does Vimeo deal with violent content?* Retrieved from https://help.vimeo.com/hc/en-us/articles/12427641451409-How-does-Vimeo-deal-with-violent-content-

Vimeo. (2023). *Report abuse and violations*. Retrieved from https://help.vimeo.com/hc/en-us/articles/12426004118417-Report-abuse-and-violations

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

Wanging, Z. (2022, March 18). *China's Instagram wants more male users. It's using women as bait.* Retrieved from https://www.sixthtone.com/news/1009886

WeChat. (2023). *We Chat - Acceptable Use Policy*. Retrieved from https://www.wechat.com/en/acceptable_use_policy.html

WeChat. (2023). *WeChat Community Guidelines*. Retrieved from https://safety.wechat.com/en_US/community-guidelines

WeChat. (2023). *WeChat Policy Implementation*. Retrieved from https://safety.wechat.com/zh_CN/enforcement/policy-enforcement

Weimann, G., & Vellante, A. (2021). The Dead Drops of Online Terrorism. *Perspectives on Terrorism , 15*(4), 39-53. Retrieved from https://www.jstor.org/stable/27044234

Weimann, G., & Vellante, A. (2021, August). *The dead drops of online terrorism: How jihadists use anonymous online platforms*. Retrieved from https://www.jstor.org/stable/27044234?seq=2

West, J. (2022, July 27). *How more robust evidence can help tackle terrorist and violent extremist content online*. Retrieved from https://www.oecd-forum.org/posts/how-more-robust-evidence-can-help-tackle-terrorist-and-violent-extremist-content-online

WhatsApp. (2021). *Terms of Service*. Retrieved from https://www.whatsapp.com/legal/terms-of-service/?lang=en

WhatsApp. (2023, How to block and report contacts). Retrieved from https://faq.whatsapp.com/1142481766359885/?helpref=hc_fnav&cms_platform=android

WhatsApp. (2023). *About account bans*. Retrieved from https://faq.whatsapp.com/465883178708358

WhatsApp. (2023, July 1). *India Monthly Report under the Information Technology Rules 2021*. Retrieved from https://scontent-cdg4-3.xx.fbcdn.net/v/t39.8562-6/356864340_258864053422020_7727900727558791209_n.pdf?_nc_cat=108&ccb=1-7&_nc_sid=ae5e01&_nc_ohc=hnuoiwlclpIAX801qOh&_nc_ht=scontent-cdg4-3.xx&oh=00_AfCN4lhi-XTh3h64qxgXwf49LpHMwmAkZczs2F6wX4MRkQ&oe=64C915BB

WhatsApp. (2023). *Seeing the message "Your phone number is banned from using WhatsApp. Contact support for help."*. Retrieved from https://faq.whatsapp.com/508673377786103/?helpref=hc_fnav

WhatsApp. (2023, July). *WhatsApp Monthly India Reports*. Retrieved from https://www.whatsapp.com/legal/india-monthly-reports

Wikimedia. (2021). *Unique devices*. Retrieved from Wikimedia Statistics: https://www.envisagedigital.co.uk/wordpress-market-share/

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

Wikimedia Foundation. (2022). *Transparency report: July to December 2022*. Retrieved from Wikimedia Foundation: https://transparency.wikimedia.org/

Wikimedia Foundation. (2023). *Universal Code of Conduct*. Retrieved from https://foundation.wikimedia.org/wiki/Policy:Universal_Code_of_Conduct

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

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

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

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

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

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

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

Wikipedia. (2023). *Moderator Tools / Automoderator*. Retrieved from https://www.mediawiki.org/wiki/Moderator_Tools/Automoderator

Wikipedia. (2023). *Responding to threats of harm*. Retrieved from https://en.wikipedia.org/wiki/Wikipedia:Responding_to_threats_of_harm

Wikipedia. (2023). *Trust and Safety*. Retrieved from https://meta.wikimedia.org/wiki/Trust_and_Safety

Wikipedia. (2023). *Trust and Safety / Case Review Committee*. Retrieved from https://meta.wikimedia.org/wiki/Trust_and_Safety/Case_Review_Committee#Submitting_appeals

Wikipedia. (2023). *Wikipedia: Administrators*. Retrieved from https://en.wikipedia.org/wiki/Wikipedia:Administrators

Wikipedia. (2023). *Wikipedia: Arbitration Committee*. Retrieved from https://en.wikipedia.org/wiki/Wikipedia:Arbitration_Committee

Wikipedia. (2023). *Wikipedia: Blocking policy*. Retrieved from https://en.wikipedia.org/wiki/Wikipedia:Blocking_policy

Wikipedia. (2023). *Wikipedia: Editing restrictions*. Retrieved from https://en.wikipedia.org/wiki/Wikipedia:Editing_restrictions

Wikipedia. (2023). *Wikipedia: General sanctions*. Retrieved from https://en.wikipedia.org/wiki/Wikipedia:General_sanctions

Wikipedia. (n.d.). *Controversial Reddit communities*. Retrieved from https://en.wikipedia.org/wiki/Controversial_Reddit_communities

Wire. (2023). *Wire Transparency Report* .

Wise, J. (2023, March). *Skype users: How many people use Skype in 2023?* Retrieved from https://earthweb.com/how-many-people-use-skype/

Wong, Q. (2023, February). *Thrown Into Facebook Jail? Meta Says It Will Explain What You Did Wrong*. Retrieved from CNET.

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

WordPress.com. (2023). *Suspended Content and Sites*. Retrieved from WordPress.com: https://en.support.wordpress.com/suspended-blogs/

X. (2023). *Notices on Twitter and what they mean*. Retrieved from https://help.twitter.com/en/rules-and-policies/notices-on-twitter#:~:text=When%20we%20permanently%20suspend%20an,believe%20we%20made%20an%20error.

X. (2023). *Our approach to policy development and enforcement philosophy*. Retrieved from https://help.twitter.com/en/rules-and-policies/enforcement-philosophy

X. (2023). *Our range of enforcement options*. Retrieved from https://help.twitter.com/en/rules-and-policies/enforcement-options

X. (2023). *The X Rules*. Retrieved from https://help.twitter.com/en/rules-and-policies/x-rules

Xiaohongshu. (2021, December). *Xiaohongshu Community Standards*. Retrieved from https://agree.xiaohongshu.com/h5/terms/ZXXY20221213003/-1

Xiaohongshu. (2023). *Community Live Streaming Standards*. Retrieved from https://agree.xiaohongshu.com/h5/terms/ZXXY20220331001/-1

Yang, Z. (2022, October). *WeChat users are begging Tencent to give their accounts back after talking about a Beijing protest*. Retrieved from https://www.technologyreview.com/2022/10/16/1061713/wechat-accounts-begging-tencent-beijing-protest/

YouTube vs. Vimeo: The key differences in 2023 (updated). (2023, May). Retrieved from https://www.lemonlight.com/blog/youtube-vs-vimeo-the-key-differences-in-2023/

YouTube/Google. (2023). *Get involved with YouTube Contributors*. Retrieved from https://support.google.com/youtube/answer/7124236?hl=en-GB&ref_topic=7124235&sjid=13166646673953227442-EU

YouTube/Google. (2023). *Taking action on violations*. Retrieved from https://www.youtube.com/intl/ALL_in/howyoutubeworks/policies/community-guidelines/#taking-action-on-violations

YouTube/Google. (2023). *YouTube Community Guidelines enforcement visible changes*. Retrieved from https://support.google.com/transparencyreport/answer/9198203?hl=en-GB

YouTube/Google. (2023). *YouTube Contributors*. Retrieved from https://contributors.youtube.com/

YouTube/Google. (2023). *YouTube's Community Guidelines*. Retrieved from https://support.google.com/youtube/answer/9288567?sjid=17414125458760940771-EU

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

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

Zhou, V. (2023, March). *Xiaohongshu is teaching young Chinese women how to buy the perfect life*. Retrieved from https://restofworld.org/2023/xiaohongshu-app-chinas-lifestyle-bible/

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

Zoom. (2023). *Acceptable Use Guidelines Enforcement*. Retrieved from https://explore.zoom.us/en/acceptable-use-guidelines-enforcement/

Zwieglinska, Z. (2023, May). *Beauty brand Klorane invests in a Twitch campaign with niche content*. Retrieved from https://www.glossy.co/beauty/beauty-brand-klorane-invests-in-a-twitch-campaign-with-niche-content/

Endnotes

[1] The fact that a particular Service appears on the Intensive Services list does not mean the OECD is saying it is a terrorist or violent extremist organisation, or that it is not such an organisation. It simply means that, according to the methodology used to create the list, the Service is used to propagate enough TVEC for it to be one of the 50 most TVEC-intensive Services.

[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, 2022).

[3] 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.

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

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

[6] 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.

[7] Dailymotion, Discord, Facebook, Facebook Messenger, Instagram, Josh, Kuaishou/Kwai, Likee, Picsart, Quora, Reddit, Snapchat, Twitch, Vimeo, WeChat, X, YouTube, Zoom

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

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

[10] 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.

[11] Bilibili, Douyin, Dropbox, Google Drive, IMO, LinkedIn, Moj, OneDrive, Pinterest, ShareChat, Skype; Teams, TikTok, Toutiao, Tumblr, Viber, Worpress.com, Xigua Video, Xiaohongshu, Youku Tudou

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

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

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

[15] Baidu Tieba, iQIYI

[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] Douban, iMessage/FaceTime, LINE, QQ, Steam, Telegram, Weibo, WhatsApp, Wikipedia

[20] According to the Anti-Defamation League, it is not clear where this document was originally posted, but Gendron reportedly claimed he planned to post it on 8chan.moe and 4chan, and send links to Discord servers.

[21] See Section 1 of the Facebook, Instagram, Snapchat, WeChat, and Dailymotion profiles.

[1] 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.

[2] 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.

[3] 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.

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

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

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

[7] Bilibili, Discord, Facebook (in Groups), Picsart, Reddit, Steam, Twitch, Viber, Wikipedia

[8] 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.

[9] 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).

[10] 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.

[11] 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).

[12] 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.

[13] See Section 5 of the profiles of Baidu Tieba, Bilibili, Dailymotion, Discord, Dropbox, Douyin, Facebook, Facebook Messenger, Google Drive, IMO, Instagram, iQIYI, Josh, Kuaishou/Kwai, Likee, LINE, LinkedIn, Moj, OneDrive, Picsart, Pinterest, QQ, QZone, Reddit, Skype, ShareChat, Snapchat, Teams, TikTok, Toutiao, Tumblr, Twitch, Viber, Vimeo, WhatsApp, WeChat, Weibo, X, Xiaohongshu, Xigua Video, YouTube, Youku Tudou, Zoom.

[14] 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

[15] 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.

[16] 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

[17] See Section 4.1 of the profiles of Baidu Tieba, Bilibili (at its discretion), Dailymotion, Douban (at its discretion), Dropbox (except if the violation engages Dropbox's legal liability), Facebook, Facebook Messenger, Google Drive (except if the violation engages Google's legal liability), Instagram, Josh (at its discretion), Kuaishou/Kwai, Likee, LinkedIn, Moj, OneDrive (at its discretion), Picsart, Pinterest (at its discretion), Quora, Reddit, Skype (at its discretion), ShareChat, Snapchat, Steam (at its discretion), TikTok, Teams (at its discretion), Toutiao, Tumblr (at discretion), Twitch, Viber (at its discretion), WeChat, Weibo (at discretion), WhatsApp, Wikipedia (for blocks), Wordpress.com (at its discretion), X, Xiaohongshu, YouTube, Zoom.

[18] 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.

[19] 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.

[20] 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

[21] See Section 4.2 of the profiles of Baidu Tieba, Dailymotion, Discord, Douyin, Dropbox, Facebook, Facebook Messenger, Google Drive, Instagram, Josh, Kuaishou/Kwai, Likee, LINE, LinkedIn, Moj, OneDrive, Pinterest, Quora, Reddit, ShareChat, Skype, Snapchat, Steam, Teams, Telegram (primarily for spam), TikTok, Toutiao, Tumblr, Twitch, Viber, Vimeo, WeChat, WhatsApp, Wikipedia (for bans), Worpress.com, X, Xiaohongshu, Xigua Video, YouTube, Zoom.

[22] Figures do not add up to 50 in this column because one of the far-right-focused services – Thedonald.win – was no longer operational, so it was not possible to determine anything about its governing documents.

[23] Mainstream TVEC-intensive Services: YouTube, X, Facebook, Instagram, TikTok and Discord.

[24] See Section 1 of the profiles of Discord, Facebook, Instagram, Justpaste.it, Rocket.Chat, Rumble, Slack, SoundCloud, Threads, TikTok, X, and YouTube.

[25] Mainstream TVEC-intensive Services: VK; File-sharing TVEC-intensive Services: Google Drive, Dropbox, Justpaste.it, Google Docs, Mega.nz, Pixeldrain; Far-right-focused TVEC-intensive Services: BitChute, Rumble, Parler, Odysee, Doxbin.

[26] See Section 1 of the profiles of DoxBin, Dropbox, Element, Google Drive, GoyimTV, Odysee, Pixeldrain, and TamTam.Chat.

[27] Mainstream TVEC-intensive Services: Telegram, WhatsApp, Element; File-sharing TVEC-intensive Services: Archive.org, Files.fm, MediaFire, File.io, Gofile.io, Anonfiles; Far-right-focused TVEC-intensive Services: Gab, Brandnewtube, Gettr, 8kun, WeGoSocial, SafeChat, Wimkin, Worldtruthvideos, Xephula.

[28] See Section 1 of the profiles of 4chan, Archive.org, ChirpWire, File.io, Gab, Gofile.io, Mastodon, Matrix, MediaFire, Signal, Telegram, Threema, WhatsApp, and Wire.

[29] File-sharing TVEC-intensive Services: Telegraph, Tlgur, Uploadgram; Far-right-focused TVEC-intensive Services: Patriots.win, Redvoicemedia.com, 88msn.com, Mzwnews.com, Thegreaterreset.org, Nordfront.dk, Lookaheadamerica.org, Patriotfront.us, Vastarinta.com, Nordicrresistancemovement.org.

[30] See section 1 of the profiles of 3pdirectory.com, Abolitionmedia.noblogs.org, Alazaimll.websites.co.in, Alqassam.ps, Americanfuturistpublishing.com, Amjaad.video, Ansarollah.com, Dalelansar.info, itarchives.com, LiveGore.com, Malhm.xyz, Moqawama.ir, Nuceciwan127.xyz, Shadanews.com, Telegraph, and Umarmediattp.org.

[31] Figures do not add up to 50 in this column because one of the far-right-focused services – Thedonald.win – was no longer operational, so it was not possible to determine anything about its approach to content moderation.

[32] Mainstream TVEC-intensive Services: Facebook, YouTube, Instagram, TikTok, X, Discord, VK; File-sharing TVEC-intensive Services: Google Drive, Dropbox, Justpaste.it, MediaFire, Google Docs, Mega.nz.

[33] See section 5 of the profiles of 4chan, Discord, Dropbox, Element, Facebook, Google Drive, Instagram, Justpaste.it, Mastodon, Matrix, MediaFire, Rocket.Chat, Soundcloud, Threads, TikTok, Wire, X, and YouTube.

[34] Mainstream TVEC-intensive Services: WhatsApp, Telegram, Element; File-sharing TVEC-intensive Services: Files.fm, Pixeldrain; Far-right-focused TVEC-intensive Services: BitChute, Rumble, Parler, Odysee, 8kun, WeGoSocial, Wimkin.

[35] See section 5 of the profiles of Odysee, Pixeldrain, Rumble, TamTam.Chat, Telegram, Threema, and WhatsApp.

[36] Far-right-focused TVEC-intensive Services: Gab, Gettr and SafeChat.

[37] See section 5 of the profiles of ChirpWire, DoxBin, File.io, Gab, Gofile.io, GoyimTV, Signal, and Wire.

[38] File-sharing TVEC-intensive Services: Archive.org and Anonfiles; Far-right-focused TVEC-intensive Services: Doxbin and Xephula.

[39] See section 5 of the profile of Archive.org.

[40] File-sharing TVEC-intensive Services: Telegraph, Tlgur, Uploadgram, File.io, Gofile; Far-right-focused TVEC-intensive Services: Patriots.win, Brandnewtube, Redvoicemedia.com, 88nsm.com, Mzwnews.com, Worldtruthvideos, Thegreaterreset.org, Nordfront.dk, Lookaheadamerica.org, Patriotfront.us, Vastarinta.com, Nordicresistancemovement.org.

[41] See section 5 of the profiles of 3pdirectory.com, Abolitionmedia.noblogs.org, Alazaimll.websites.co.in, Alqassam.ps, Americanfuturistpublishing.com, Amjaad.video, Ansarollah.com, Dalelansar.info, itcarchive.org, LiveGore.com, Malhm.xyz, Moqawama.ir, Nuceciwan127.xyz, Shahadanews.org, Telegraph, and Umarmediattp.org.

[42] Mainstream TVEC-intensive Services: Facebook, YouTube, WhatsApp, Instagram, TikTok, X, VK ; File-sharing TVEC-intensive Services: Google Drive, Dropbox, Google Docs, MediaFire, Mega.nz; Far-right-focused TVEC-intensive Services: Bitchute, Gettr (discretional), SafeChat (discretional).

[43] See Section 4.1 of the profiles of Dropbox, Google Drive, Mastodon, MediaFire, Rocket.Chat (for law enforcement requests), Soundcloud, Threads, TikTok, WhatsApp, X, and YouTube.

[44] Mainstream TVEC-intensive Services: Facebook, YouTube, Instagram, WhatsApp, TikTok, X, Discord, VK; File-sharing TVEC-intensive Services: Google Drive, Dropbox, Google Docs, MediaFire, Mega.nz; Far-right-focused TVEC-intensive Services: Bitchute, Gab, Gettr (discretional), SafeChat (discretional).

[45] See Section 4.2 of the profiles of 4chan, Discord, Dropbox, Google Drive, Justpaste.it, Mastodon, MediaFire, Odysee, Rocket.Chat, SoundCloud, Telegram (primarily for spam), TikTok, Threads, WhatsApp, X, and YouTube.

[46] Mainstream TVEC-intensive Services: Telegram, Element; File-sharing TVEC-intensive Services: Telegraph, Archive.org, Files.fm, Tlgur, Pixeldrain, Uploadgram, File.io, Gofile.io, Anonfiles; Far-right-focused TVEC-intensive Services: 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, Nordicresistancemovement.org.

[47] See Sections 4.1 and 4.2 of the profiles of 3pdirectory.com, Abolitionmedia.noblogs.org, Alazaimll.websites.co.in, Alqassam.ps, Americanfuturistpublishing.com, Amjaad.video, Ansarollah.com, Archive.org, ChirpWire, Dalelansar.info, DoxBin, Gab, Gofile.io, Element, File.io, GoyimTV, itcarchive.org, LiveGore.com, Matrix, Malhm.xyz, Moqawama.ir, Nuceciwan127.xyz, Pixeldrain, Rumble, Shahadanews.org, Signal, Slack, TamTam.Chat, Threema, Telegraph, Umarmediattp.org, and Wire.

[48] Available at https://www.canada.ca/en/canadian-heritage/campaigns/harmful-online-content.html

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

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

[51] See subsection “Disclosures by Chinese Platforms” in Section 3 of the Report.

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

[53] 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.

[54] 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

[55] Tumblr Live was first introduced in the United States only. As of August 2023, it is now also available in Brazil, Canada, the European Union, Japan, Malaysia, Mexico, Korea, Türkiye, and the United Kingdom.

[56] The fact that a particular Service appears on the Intensive Services list does not mean the OECD is saying it is a terrorist or violent extremist organisation, or that it is not such an organisation. It simply means that, according to the methodology used to create the list, the Service is used to propagate enough TVEC for it to be one of the 50 most TVEC-intensive Services.