Organisation for Economic Co-operation and Development

DSTI/CDEP(2021)7/FINAL

Unclassified **English - Or. English**

26 January 2022

DIRECTORATE FOR SCIENCE, TECHNOLOGY AND INNOVATION
COMMITTEE ON DIGITAL ECONOMY POLICY

Voluntary Transparency Reporting Framework for TVEC, version 1.0

JT03488676

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

Note by the Secretariat

This is the first edition of the Voluntary Transparency Reporting Framework (VTRF) for terrorist and violent extremist content (TVEC) online, which was approved and declassified by CDEP at its 85th Session on 2 December 2022. VTRF 1.0 has been developed in consultation with a multi-stakeholder experts' group in which more than 100 of the world's leading figures on platform governance and/or violent extremism and terrorism activity online have participated. With its background information, questionnaire and glossary of key terms, the VTRF aims to provide a common standard for a baseline level of transparency reporting on TVEC. The VTRF was designed for use by any online content-sharing service, regardless of its business model, size, reporting experience or approach to content moderation.

The VTRF will next take the form of a convenient web portal where companies can voluntarily submit TVEC transparency reports by completing the questionnaire online, and where policy makers, researchers, other services, and anyone else can view the submitted information. A pilot version of the portal is expected to launch in March 2022 at oecd-vtrf-pilot.org.

The OECD's work on TVEC online is proceeding with the kind financial support of Australia, Korea and New Zealand. The Secretariat would also like to thank all members of the TVEC Experts' Group for their contributions to this work.

1. Background

Origins and Aims of the VTRF

A free, open and secure Internet is a powerful tool to promote connectivity, enhance social inclusiveness and foster economic growth. The Internet is not, however, immune from abuse by terrorist and violent extremist actors. Terrorist and violent extremist content (TVEC) has adverse impacts on human rights, security and society.

The Voluntary Transparency Reporting Framework (VTRF) project began in 2019 as part of the OECD's response to a number of initiatives, statements and international calls for action that emerged that year, all seeking to address this challenge. The G20 Osaka Leaders' Statement on Preventing Exploitation of the Internet for Terrorism and Violent Extremism, the G7 Digital Ministers Chair's Summary, the G7 Interior Minister outcome document Combating the Use of the Internet for Terrorist and Violent Extremist Purposes, and the communiqué from the Joint meeting of Five Country Ministerial and quintet of Attorneys-General (London 2019) reflect countries' commitments to work together to counter the use of the Internet for terrorism and violent extremism without compromising human rights and fundamental freedoms. Furthermore, the Christchurch Call to Eliminate Terrorist and Violent Extremist Content Online contains commitments to act by both governments and online content-sharing service providers.

One of the common themes in the calls for action on TVEC is the need for greater transparency from online content-sharing services regarding their community standards or terms of service, including descriptions of their policies and procedures for detecting and actioning TVEC as well as information on how they actually address situations in which TVEC is shared. The companies that formally support the Christchurch Call, for example, committed to implement regular and transparent public reporting, in a way that is measurable and supported by clear methodology, on the quantity and nature of TVEC being detected and removed.

However, relatively few online content-sharing services currently provide transparency reports specifically on terrorist and/or violent extremist content (OECD, 2020[1]; OECD, 2021[2]). Even within this group, there is great variation in approaches and levels of transparency. Many online content-sharing services, particularly smaller services, do not presently provide transparency reports, or similar types of information, at all. That may be because it is challenging for them to devote the necessary resources required to issue such reports. In any event, these conditions create challenges for stakeholders seeking to gain a holistic understanding of the measures taken by the full ecosystem of online content-sharing services to address TVEC and the consequences of those measures for human rights, not just for individual persons but also for larger groups (Human Rights Watch, 2020[3]; Abdul Rahman Al Jaloud, 2019[4]).

At the 2019 G7 meeting in Biarritz, Australia's Prime Minister announced that Australia was working with New Zealand to fund a new project at the OECD to improve voluntary transparency reporting on terrorist and violent extremist content (TVEC) by online content-sharing services. Since then, Canada and Korea have provided further resources for the work.

The VTRF is the heart of the project. It aims to provide a common standard for TVEC transparency reporting through a questionnaire that any online content-sharing service can use to provide information about its TVEC-related policies and actions, no matter what its business model, size, reporting experience and approach to content moderation are. The VTRF also contains a glossary of key terms (such as “content-sharing service”) relevant to TVEC transparency reporting.

To inform the VTRF’s development, the OECD convened an international, multidisciplinary advisory body consisting of experts from government, the tech sector, civil society and academia. More than 100 experts contributed to this effort. Their expertise encompasses counter-terrorism, counter-violent extremism, digital governance and policy, law, civil liberties, and human rights including the freedom of expression. The group includes leading figures who are recognised as authorities on addressing violent extremist and terrorist activity online and/or in platform governance. The work of the TVEC experts’ group is overseen by the OECD’s Committee on Digital Economy Policy.

The VTRF is an opportunity to take demonstrable action in response to the calls for greater transparency. It also provides a potential avenue for reducing the risk of regulatory fragmentation while complementing existing initiatives. It aims to strengthen the evidence base on TVEC by providing a flexible but robust measurement framework. It is meant to encourage more online content-sharing services to develop transparency reports, and to build the level of understanding of other stakeholders on the measures that online content-sharing services take to combat TVEC. It is also meant to raise understanding on the services’ appeal and redress mechanisms. The VTRF further aims to support international cooperation and information sharing, and ultimately to reduce the volume and reach of TVEC online while promoting the protection of human rights. In addition, by providing a voluntary reporting standard that potentially all OECD countries, and perhaps more, can support, the VTRF aims to promote coordinated national standards going forward. It therefore has the potential to decrease burdens involved with transparency reporting, broaden the scope for effective international dialogue, comparisons and research, and promote a common and more coherent approach to transparency reporting on terrorist and violent extremist content and activity.

This initial edition of the VTRF, or VTRF 1.0, asks only for information that is considered appropriate for a baseline level of transparency reporting. That is to say, all of the top-level questions are intended to be answerable by any online content-sharing service. Many of the sub-questions are optional or are asked only if it is clear that the respondent can answer them. Future editions of the framework will have additional categories to recognise transparency that goes above and beyond the baseline level. The criteria for getting into those categories may be scaled so that they are fair and proportionate to the capabilities of the respondent. In other words, smaller services with fewer resources would be recognised for above-baseline reporting without having to meet all of the same reporting standards as very large, well-resourced services. The purpose of having additional categories is to motivate more comprehensive and innovative reporting by recognising efforts to report more, and more detailed, metrics, as well as efforts to experiment with new or additional technologies, metrics and/or definitions (or to refine existing ones).

The VTRF recognises that online content-sharing services need the flexibility to innovate and develop their approaches, particularly in view of their starting points, capacity, and the type of services they offer, as the environment changes. As the nature of the threat posed by terrorism and violent extremism evolves, the OECD, again in consultation with a multi-stakeholder group of experts, will need to review and revise the VTRF periodically to ensure that it remains an up-to-date standard.

Understanding the Term “Terrorist and Violent Extremist Content”

The VTRF does not impose a definition of TVEC. Instead, it provides a common way for content-sharing services to report while working with the definitions of their own choosing, and to improve upon them over time, with the goal that the services will converge on the best definition(s) over time.

Presently, there are no universally accepted definitions of either terrorism or violent extremism, nor by extension, of terrorist and violent extremist content, which can present challenges for all actors involved in addressing such content, including online content-sharing services. While some online content-sharing services have developed and published their own definitions and understandings, many others have not (OECD, 2020[1]) (OECD, 2021[2]). In the absence of universally agreed definitions, different approaches have been taken which seek to clarify what falls within the scope of the terms. That different approaches have been taken reflects the fact that no single approach has yet garnered universal consensus, and that there are advantages and disadvantages associated with different approaches.

A number of lists of organisations and individuals designated as terrorist(s) or violent extremist(s) have been published by, among others, the United Nations Security Council, national governments, academic institutions and civil society organisations. Such lists can provide a useful starting point for companies developing policies relating to TVEC on their platforms, and many online content-sharing services currently make use of such lists.

There are, however, limitations and challenges that accompany the use of such lists. Lists are less likely to capture lone perpetrators of terrorism or violent extremism who may be influenced by and/or produce terrorist or violent extremist content online. There are also challenges in developing a single comprehensive list that is global, updated in real time and free from any bias (Global Research Network on Terrorism and Technology, 2019[5]). An approach based on lists also means that content is removed on the basis of the author or speaker, rather than its substance, which raises challenges from a freedom of expression perspective. Finally, approaches based exclusively on lists risk silencing a part of a human conflict and thereby affecting the right to memory and truth that is recognised and protected as a fundamental right in some regions of the world (Inter-American Commission on Human Rights, 2019[6]).

In addition to these lists, there have been a number of efforts to provide guidance on approaches to understanding terrorism and violent extremism, particularly the former, that may also be helpful for developing approaches to understanding TVEC based on its content, rather than on the speaker alone (sometimes known as "content-based approaches"). These efforts focus not on organisations and individuals, but on the nature of the act or content itself. Some online content-sharing services take such an approach to understanding TVEC on their platforms and a number of resources provide guidance for companies seeking to take such an approach (Inter-American Commission on Human Rights, 2019[6]).

While a content-based approach may capture a broader range of content than a list-based approach, no single approach or proposed definition has yet been agreed by governments at the international level, and definitions in national legislation continue to vary. Content-based approaches also raise questions around how to ensure that the right to freedom of expression is not adversely impacted, although guidance on this question has been developed by the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism (United Nations Human Rights Council, 2010[7]). Content-based approaches also face important challenges in bringing the rights and duties of truth and memory to the digital environment (Inter-American Commission on Human Rights, 2019[6]).

A number of online content-sharing services combine both approaches, making decisions on the basis of particular lists and a qualitative assessment of content. The services may also remove content that could be considered TVEC on the basis of other forms of content prohibited by their policies and terms of service, such as criminal activity, dangerous organisations, graphic violence or hate speech.

Given the challenges involved in reaching consensus on the terms "terrorism" and "violent extremism", and the many existing forums and processes whose work encompasses the issue, the VTRF does not attempt or seek to define either term. While it notes some of the approaches that companies may wish to consider, the VTRF does not propose or endorse any definitions or particular approaches, nor does it require or expect companies who use the VTRF to adopt any particular definitions of the terms or approaches.

The VTRF does, however, seek to enhance the level of transparency over whether and how companies define these terms, bearing in mind both the evolving nature of the threat posed by terrorist and violent extremist actors and the risks to human rights that might emerge in potential responses to this threat. Through setting out questions and metrics that will lead to this transparency, the VTRF also aims to help companies new to transparency reporting think through the challenges involved when it comes to their understanding of the terms, as well as to provide an opportunity for companies to learn from the practices of others.

The VTRF Is Intended to Be Fully Consistent with the Rule of Law and Respect for Human Rights

Nothing in the VTRF should be understood as requiring or encouraging companies to do anything that would be inconsistent with the rule of law, respect for human rights and fundamental freedoms, or compliance with international law and the national law within the jurisdictions within which they operate.

2. Instructions for Completing the Questionnaire and Interpreting the Responses

The VTRF is intended to be used by any online content-sharing service (a term defined in the Glossary) that wishes to use it, no matter what its business model, size, reporting experience and approach to content moderation are. Nevertheless, the VTRF is designed to focus on terrorist and violent extremist content. If such content is never posted, stored or shared on a particular service, then there may not be much reason for that service to complete the VTRF questionnaire, unless it is simply to show that the service has relevant policies and procedures in place anyway.

The VTRF questionnaire is a series of qualitative and quantitative questions. Many are in a simple Y/N or multiple choice format and can be answered with a click of the mouse. Others call for data, a more descriptive response, or both.

Virtually every question provides an opportunity to add clarifications or additional information if the respondent wishes to do so. Doing so is entirely optional. In fact, it bears repeating that the "V" in VTRF stands for "voluntary". Completing this questionnaire is a voluntary undertaking.

Some questions are nested within others. However, there is no expectation that every respondent will answer every nested sub-question. That is because the VTRF is designed to ask certain questions only if they are relevant to the respondent. If they are not relevant, then the VTRF will guide the respondent past them. This is why the questionnaire currently has text at the start of some questions such as "If you answered yes to Question X, then . . ." It is there only to help readers understand the intended flow of the questions. When implemented, the VTRF questionnaire will take the form of an online decision tree, where such text will no longer be needed because it will already be taken into account by the decision tree software. Thus, the text will be shorter and simpler, and respondents will not even see certain sub-questions unless they are triggered by a previous response indicating that they will be pertinent.

The glossary in Annex A clarifies a number of terms and may be helpful in case the intended meaning of a term used in the questionnaire seems unclear.

The questionnaire is intended to be completed at the individual online content-sharing service, or "platform" level, not necessarily at the corporate-wide level. Therefore, entities that own more than one content-sharing service are asked to complete a separate questionnaire for each service. However, if an individual service features multiple functionalities, it is not necessary to complete a separate questionnaire for each functionality. For example, if a company owns a social media platform and a separate video streaming platform, it should complete two questionnaires. But if the social media platform also has functionality for, say, livestreaming, it is unnecessary to fill out two separate questionnaires for that one platform.

It is recognised that different jurisdictions have different definitions of terrorism and violent extremism. VTRF 1.0 asks for global totals, not geographic breakdowns. Therefore, even though definitions change from country to country and that affects the amount of content flagged, for example, by government Internet Referral Units, all of that flagged content should be counted by the respondent.

The questionnaire's purpose is to illuminate what is rather than to suggest what should be. In that vein, "success" in completing the questionnaire depends on whether the questions are answered, not on the nature of the answers themselves. In some cases the answer may simply be "No, we do not have that data". As far as VTRF 1.0 is concerned, that is a valid answer. In other words, although respondents are encouraged to provide as much data as they can, an inability to provide data should not reflect negatively on them.

It is also recognised that online content-sharing services can provide multiple services and use a variety of strategies for detecting and actioning harmful content, including content that can be considered terrorist and/or violent extremist content. For example, some rely primarily on users and/or trusted notifiers to detect such content; others also deploy automated technologies. Likewise, some cast a wide net and eventually action a small proportion of flagged content; others use a narrower approach and action a higher percentage. Furthermore, using multiple detection methods can mean that some content is flagged or reported several times but actioned only once. Therefore, certain responses (or lack thereof) to the questions could, if considered out of context, lead to a misimpression of under-action, over-action or low accuracy. To avoid that and other misinterpretations, respondents and readers should understand that responses (or lack thereof) to the questions do not necessarily imply anything negative about the detection or moderation strategies or practices of the online content-sharing services submitting the responses."

Finally, because the VTRF is new and will necessarily evolve, the implementation of version 1.0 will be a pilot programme that gathers user feedback. That feedback will enable improvements in version 2.0, on which work is expected to begin in late 2022. The questionnaire will therefore be followed by a short feedback survey. All visitors to the VTRF web portal will be asked to provide feedback.

3. Questionnaire

1. Which of the following best describes the primary type(s) of online service(s) that you provide? (Click all that apply.)

□	Blogging
□	Cloud-based storage and sharing
□	Gaming
□	Image/message board
□	Livestreaming
□	Messaging
□	Social media
□	Video chat
□	Video sharing
□	Video streaming
□	Other (if selected, please describe)

If you would like to add a comment or other information, please do so here.

2. Do you prohibit terrorist and/or violent extremist content on your service?

□	Yes
□	No

(If no, the respondent is taken to the open-ended question at the very end of Q2, after which the questionnaire ends.)

If yes:

➢ Do you use one or more specific, publicly available definitions or understandings of terrorist and/or violent extremist content?

□	Yes
□	No

- ► If yes, please provide the publicly available definition(s) or understanding(s) that you use, along with your relevant terms of service or policies.
  If you use one or more list-based definitions (such as the United Nations Security Council Consolidated List) but do not wish to reveal a particular list (or lists) you use, please note that there is no obligation to do so. However, in that case, please disclose that you use a list (or lists).

- ► If no, do you consider terrorist and/or violent extremist content to be included in other categories of content that you prohibit on your service, such as criminal activity, incitement to violence, graphic violence or hate speech?

□	Yes
□	No

- → If yes, please provide the publicly available definition(s) or understanding(s) that you use for those categories of content, along with your relevant terms of service or policies.
- → (If no, the questionnaire ends after the open-ended question at the very end of Q2.)

If you would like to add other comments or information, please do so here.

3. Do you use any of the following methods to **detect** terrorist and/or violent extremist content on your platform or service? Click on each method that you use.

□	A. Flagging by individual users or entities
□	B. Trusted notifiers
□	C. Government referrals (these are based on possible violations of your terrorist and/or violent extremist content policies and are not legally enforceable, unlike government legal requirements)
□	D. Government legal requirements (these are legally enforceable)
□	E. Internal flagging done by human review
□	F. Internal flagging done by automated technologies
□	G. Hybrid system (technological and human detection)
□	H. Cross-company shared databases or tooling
□	I. Other (if selected, please describe)

If you selected any of the preceding choices and you answered yes to the Q2 sub-question “Do you use one or more specific, publicly available definitions or understandings of terrorist and/or violent extremist content?”:

- ➢ Can you determine the total amount of content that was flagged or reported as terrorist and/or violent extremist content on your service during the reporting period?

□	Yes
□	No

If yes:

- Are you willing and able to disclose it?

□	Yes
□	No

  If yes, how much content, in total, was flagged or reported as terrorist and/or violent extremist content on your service during the reporting period? Please provide only a quantitative answer here. You may use the open text box at the end of Question 3 to explain your detection strategy as well as your methodology for answering this sub-question.

  If no, please explain why.

- Can you determine the amounts of content that are flagged or reported as terrorist content separately from the amounts of content that are flagged or reported as violent extremist content on your service?

□	Yes
□	No

  If yes:

  - Are you willing and able to disclose them?

□	Yes
□	No

    - If yes, please provide a breakdown of the total amount of content flagged or reported as terrorist and violent extremist content, respectively, on your service during the reporting period. Please provide only a quantitative answer here. You may explain your methodology for answering this sub-question, indicating what you included in your data, in the next sub-question.
    - If no, please explain why.

If you selected any of the detection methods (A – I) above, you answered no to the Q2 sub-question "Do you use one or more specific, publicly available definitions or understandings of terrorist and/or violent extremist content?", and you answered yes to the further sub-question "do you consider terrorist and/or violent extremist content to be included in other categories of content that you prohibit on your service, such as criminal activity, incitement to violence, graphic violence or hate speech?":

➢ Can you determine the total amount of content that was flagged or reported on your service under the categories in which you consider terrorist and/or violent extremist content to be included, during the reporting period?

☐	Yes
☐	No

If yes:

► Are you willing and able to disclose it?

☐	Yes
☐	No

- If yes, how much content, in total, was flagged or reported on your service under the categories in which you consider terrorist and/or violent extremist content to be included, during the reporting period? Please provide only a quantitative answer here. You may use the open text box at the end of Question 3 to explain your detection strategy as well as your methodology for answering this sub-question.

- If you are able to break that number down into discrete categories, please do so. For example, if you can disaggregate it into the various categories of prohibited content in which it belongs, according to your policies, please do. You can explain your methodology, indicating what you included in your data, in the next sub-question.

- If no, please explain why.

If you would like to add other comments or information, please do so here. For example, if you use shared cross-company databases or tooling, you can provide examples here. If applicable, this is also where you can explain your detection strategy and your methodology for answering any quantitative sub-questions, indicating what you included in your data. A particular concern is double counting, which can occur when the same piece of content is flagged by multiple sources/methods. There is no way to avoid such double counting unless your service counts just the flagged *content* itself, as opposed to the number of *times* each piece of flagged content was flagged. For example, the same piece of content might be flagged by several referral units, several trusted flaggers, and automated technologies, too. Do you report that as 1 or a number greater than 1? The VTRF does not prescribe a method; it simply asks that you explain your approach here.

4. If you replied yes to the sub-question under question 3 "Can you determine the total amount of content that was flagged or reported as terrorist and/or violent extremist content on your service during the reporting period?", or you replied yes to the sub-question under question 3 "Can you determine the total amount of content that was flagged or reported on your service under the

categories in which you consider terrorist and/or violent extremist content to be included, during the reporting period?”,

Can you determine the total amount of content that is flagged or reported as a) terrorist and/or violent extremist content (if you keep track of it that way), or b) content belonging in the categories in which you consider terrorist and/or violent extremist content to be included (if you do not have specific categories for terrorist and/or violent extremist content), **according to the method of detection**?

□	Yes
□	No

If you would like to add any comments or you can provide any relevant data, please do so here. It would be particularly helpful for you to indicate, if possible, the amount of content detected for each detection method you use, at least for government referrals and government legal requirements if not for all methods. As a reminder, possible detection methods include:

- flagging by individual users or entities
- trusted notifiers
- government referrals
- governments legal requirements
- internal flagging by humans
- internal flagging by automated technologies
- hybrid systems (technological and human detection)
- cross-company shared databases or tooling
- other methods (please specify).

If that data can be further broken down by content type (such as terrorist content versus violent extremist content, if you keep track of it that way), please do so.

5. If you selected any of the detection methods in question 3, please select all interim or final **actioning** methods that you use on terrorist and/or violent extremist content:

□	a. Content removal
□	b. Content warning labels
□	c. Content hiding/quarantine
□	d. Content blocking
□	e. Warnings to account holder
□	f. Suspension/removal of account
□	g. Limitations to account functions and/or tooling
□	h. Other (if you select h, please describe)

If you would like to add other comments or information, please do so here. Please note that you will be asked to provide quantitative information about actioning in Question 6, so there is no need to include it here.

6. If you action **content** (i.e. if you ticked any of the boxes a through d in Q5) and you answered yes to the Q2 sub-question "Do you use one or more specific, publicly available definitions or understandings of terrorist and/or violent extremist content?"

Can you determine the total amount of terrorist and/or violent extremist content on which you took action during the reporting period?

□	Yes
□	No

If yes:

► Are you willing and able to disclose it?

□	Yes
□	No

- If yes, please provide that amount, along with any breakdowns that are available. For example, if you can disaggregate actioned content according to the method of detection and/or the actioning method, and/or you can separate the amount of terrorist content actioned from the amount of violent extremist content actioned, please do. You will be able to explain your methodology in the final sub-question. As a reminder, possible detection methods include

  - flagging by individual users or entities
  - trusted notifiers
  - government referrals
  - governments legal requirements
  - internal flagging by humans
  - internal flagging by automated technologies
  - hybrid systems (technological and human detection)
  - cross-company shared databases or tooling
  - other methods (please specify)

  and possible methods for actioning terrorist and/or violent extremist content include

  - content removal
  - content warning labels
  - content hiding/quarantine

- content blocking
- other method(s) (please specify)

- If no, please explain why.

If you action **content** (i.e. if you ticked any of the boxes a through d in Q5), you answered no to the Q2 sub-question "Do you use one or more specific, publicly available definitions or understandings of terrorist and/or violent extremist content?", and you answered yes to the further sub-question "do you consider terrorist and/or violent extremist content to be included in other categories of content that you prohibit on your service, such as criminal activity, incitement to violence, graphic violence or hate speech?"

Can you determine the total amount of content in the categories in which terrorist and/or violent extremist content is included, according to your policies, that you actioned during the reporting period?

☐	Yes
☐	No

If yes:

► Are you willing and able to disclose it?

☐	Yes
☐	No

➢

- If yes, please provide that amount, along with any breakdowns that are available. For example, if you can disaggregate actioned content according to the method of detection and/or the actioning method, and/or you can disaggregate it into the various categories of prohibited content in which it belongs, according to your policies, please do. You will be able to explain your methodology in the final sub-question. As a reminder, possible detection methods include

  - flagging by individual users or entities
  - trusted notifiers
  - government referrals
  - governments legal requirements
  - internal flagging by humans
  - internal flagging by automated technologies
  - hybrid systems (technological and human detection)
  - cross-company shared databases or tooling
  - other methods (please specify)

and possible methods for actioning terrorist and/or violent extremist content include

- content removal
- content warning labels
- content hiding/quarantine
- content blocking
- other method(s) (please specify)

- If no, please explain why.

If you action **accounts** (i.e. if you ticked any of the boxes e through g in Q5) and you answered yes to the Q2 sub-question “Do you use one or more specific, publicly available definitions or understandings of terrorist and/or violent extremist content?”,

Can you determine the total number of accounts on which you took action during the reporting period for violations of your policies against the use of your service for terrorist and/or violent extremist purposes as a percentage of the average number of monthly active accounts during the reporting period?

□	Yes
□	No

If yes:

► Are you willing and able to disclose it?

□	Yes
□	No

- If yes, please provide that percentage, along with any breakdowns that are available. For example, if you can disaggregate the percentage of actioned accounts according to the method of detection and/or the actioning method, and/or you can separate the portion actioned for violations of your policies against terrorism from the portion actioned for violations of your policies against violent extremism, please do. You will have the opportunity to explain your methodology in the final sub-question. As a reminder, possible detection methods include

  - flagging by individual users or entities
  - trusted notifiers
  - government referrals
  - governments legal requirements
  - internal flagging by humans
  - internal flagging by automated technologies
  - hybrid systems (technological and human detection)

- cross-company shared databases or tooling
- other methods (please specify)

and possible methods for actioning accounts include

- warnings
- account suspension/removal
- limitations to account functions and/or tooling
- other method(s) (please specify)

- If no, please explain why.

If you action **accounts** (i.e. if you ticked any of the boxes e through g in Q5), you answered no to the Q2 sub-question “Do you use one or more specific, publicly available definitions or understandings of terrorist and/or violent extremist content?”, and you answered yes to the further sub-question “do you consider terrorist and/or violent extremist content to be included in other categories of content that you prohibit on your service, such as criminal activity, incitement to violence, graphic violence or hate speech?”,

Can you determine the total number of accounts on which you took action during the reporting period for violations of those categories of your policies, as a percentage of the average number of monthly active accounts during the reporting period?

□	Yes
□	No

If yes:

- ► Are you willing and able to disclose it?

□	Yes
□	No

- If yes, please provide the percentage, along with any breakdowns that are available. For example, if you can disaggregate the percentage of actioned accounts according to the method of detection and/or the actioning method, and/or you can separate the percentages actioned for violations of each of the categories of prohibited content in which terrorist and/or violent extremist is included, according to your policies, please do. You will be able to explain your methodology in the final sub-question. As a reminder, possible detection methods include

  - flagging by individual users or entities
  - trusted notifiers
  - government referrals
  - governments legal requirements

- internal flagging by humans
- internal flagging by automated technologies
- hybrid systems (technological and human detection)
- cross-company shared databases or tooling
- other methods (please specify)

and possible methods for actioning accounts include

- warnings
- account suspension/removal
- limitations to account functions and/or tooling
- other method(s) (please specify)

- If no, please explain why.

If you would like to add other comments or information, please do so here. If applicable, this is where you can explain your methodology for providing any quantitative information about actioning content and/or accounts.

7. If your service includes livestreaming functionality (even if it is not among what you consider to be the primary functionalities), then given the potential for terrorists and violent extremists to exploit livestreaming in ways that could promote, cause, or publicize imminent violence or physical harm, do you implement controls or proactive risk parameters (such as, but not limited to, delays; restrictions for those who have previously violated your terms of service, community guidelines, or community standards; minimum audience requirements; or priority flags for livestreamed content) on livestreaming to reduce misuse?

☐	Yes
☐	No

If yes:

► Please describe the controls or proactive risk parameters that you implement, as well as the enforcement actions that you take when users are found to violate your livestreaming policies. If you can provide any accompanying data on enforcement actions, please do.

If you would like to add other comments or information, please do so here.

8. Please provide details on how you balance the need to action terrorist and/or violent extremist content with the risk that such content can be mislabelled and may actually be denouncing and documenting human rights abuses, or that it does not otherwise violate your terms of service. For example, this balancing could be carried out with measures such as independent reviews and audits of the decisions made (this refers to reviews carried out as a matter of course, rather than

to appeals) or the restoration of content to ensure that your controls, proactive risk parameters and other actioning methods for terrorist and/or violent extremist content are not misused.

9. Do you have an appeal or redress process for content and/or account actioning decisions made under your terms of service on terrorist and/or violent extremist content?

□	Yes
□	No

If yes:

- Please provide a detailed overview of those processes.

- Is your **appeal or redress process** available to the user who posted the content or owns the account in question?

□	Yes
□	No

  - If yes, Is the **outcome** of your appeal and redress process available to the user who posted the content or owns the account in question?

□	Yes
□	No

  - If no, but you would like to add an explanation or other comments, please do so here.

- Is your **appeal or redress process** available to the person or entity who requested actioning?

□	Yes
□	No

  - If yes, is the **outcome** of your appeal and redress process available to the person or entity who requested actioning?

□	Yes
□	No

    - If no, but you would like to add an explanation or other comments, please do so here.

  - ➢ If no, but you would like to add an explanation or other comments, please do so here.

- ➢ What is the total number of appeals received from all sources, during the reporting period, following content or account actioning decisions under your policies against terrorist and/or violent extremist content?

  - ► How many such appeals were decided during this reporting period (regardless of when those appeals were received)?

  - ► Of those, how many were granted?

  - ► If you can break these numbers (appeals received, decided and granted) down with any more detail, please do so here. For example, if you action both content and accounts and you are able to disaggregate the figures you just provided according to whether they concern appeals of content decisions versus account decisions, please do so here. Other possible breakdowns include: 1) by source of appeal (i.e. who appealed, a user who posted the content or owns the account in question, or the person or entity who requested actioning?); 2) by type of policy violation alleged (i.e. appeals of decisions under your policies against terrorist content versus those under your policies against violent extremism, if you track them that way; alternatively, if you consider terrorist and/or violent extremist content to be included in other categories of content that you prohibit on your service, such as criminal activity, dangerous organisations, incitement to violence, graphic violence or hate speech, then please disaggregate the number that way); 3) by method of first detection (for content actioning decisions).

If you would like to add other comments or information, please do so here.

10. How, and how often, do you measure and evaluate the efficacy and/or room for improvement of your policies in each of the following areas? Please describe, also noting the factors that prompt you to review the policies.

    - ➢ defining terrorist and/or violent extremist content
    - ➢ detecting terrorist and/or violent extremist content
    - ➢ actioning terrorist and/or violent extremist content
    - ➢ appeal/redress processes

11. Do you have a point of contact (such as a dedicated email alias, desk or department) that can be contacted during a real-world or viral event with direct online implications and which works to address harmful content on your service?

☐	Yes
☐	No

- ➢ If yes, provide any additional details that you would like to add about this point of contact. Please note that you are not being asked to provide the actual contact information here.

If you would like to add other comments or information, please do so here.

12. Are you a member of an international crisis protocol aimed at reducing the volume and impact of terrorist and/or violent extremist content online during a crisis?

□	Yes
□	No

If yes:

- ➢ Please identify the protocol.
- ➢ Did your company participate in or benefit from the activation of a crisis protocol aimed at reducing the volume and impact of terrorist and/or violent extremist content during the reporting period?

□	Yes
□	No
□	No such crisis protocol was activated during the reporting period.

- ➢ If yes, please identify the protocol.
- ➢ If no, please explain why.

If no:

Do you cooperate in any other ways with industry or law enforcement during a crisis situation involving a content incident?

□	Yes
□	No

- ➢ If yes, please explain.

If you would like to add other comments or information, please do so here.

13. Have your policies or definitions for terrorist and/or violent extremist content changed since you last completed a VTRF questionnaire in a way that affected your responses in this one?

□	Yes
□	No

- If yes, please describe the relevant changes and how they have affected your responses. This will facilitate comparisons between the quantitative information in your questionnaire responses from one period to another.

If you would like to add other comments or information, please do so here.

A Note on Future Editions of the VTRF

This section contains brief overviews of a few themes that are likely to appear in future editions of the VTRF, which will feature a standard (or "baseline") category of reporting similar to what appears in this edition, plus one or two additional categories for reporting more comprehensive or sophisticated information if the respondent wishes to do so. The themes described below would be considered beyond the standard category of reporting. They are mentioned here to provide advance insights into the VTRF's expected evolution, as well as to give companies that are already issuing transparency reports on terrorist and/or violent extremist content some ideas about what else they could usefully include now, should they wish to report more information than is requested in the questions above. It is left up to the reporting companies to decide exactly which additional metrics, if any, they wish to report, but the thematic overviews provided here may generate some ideas. Supplemental information can be included in the space below this box. Any such information provided may help to inform the development of future editions of the VTRF.

1. Processes for government referrals and government legal requirements

Future versions of the VTRF may seek additional information about your processes for government referrals and government legal requirements related to terrorist and/or violent extremist content, based on your service's policy. The purpose will be to help identify best practices for receiving and addressing such referrals and legal requirements, as well as communicating this information to the public and relevant stakeholders, while respecting human rights.

2. Management of online crises related to terrorist and/or violent extremist content

Future versions of the VTRF may ask you to provide details on your online crisis management policy with respect to terrorist and/or violent extremist content, including the geographies and languages covered by your relevant crisis response team and the consideration given to different cultural contexts. The purpose will be to understand more about the capabilities and breadth of your online crisis management programme with respect to terrorist and/or violent extremist content, as well as establish best practices for communicating this information to the public and relevant stakeholders.

3. Time required to address content and appeals

Future editions of the VTRF may ask for quantitative information about the time required to moderate content flagged as terrorist and/or violent extremist content by various sources of detection, as well as the time required to review appeals. While recognizing that faster is not always necessarily better, the purpose will be to better gauge the speed with which terrorist and/or violent extremist content flags and appeals are addressed, as well as establish best practices for communicating this information to the public and relevant stakeholders.

4. Prevalence of user interaction with terrorist and/or violent extremist content

Future editions may request information about how often terrorist and/or violent extremist content was viewed, shared, linked, commented on, and reacted to prior to being actioned, and/or about the proportion of such interactions in comparison with content interactions overall. The purpose will be to better understand terrorist and/or violent extremist content's reach and impact.

Space for Optionally Reporting Additional Metrics

If you wish to report any additional metrics or information, please use this space to do so.

4. User Feedback Survey

1. How can we improve the VTRF’s **content?**
2. How can we improve the VTRF’s **format?**
3. Do you have any additional comments or feedback on the VTRF?

Annex A. Glossary

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

> Banning – Banning a user prohibits them from logging on to a content-sharing service and/or from creating and using any new accounts.
>
> Disabling/de-activating/suspending – Disabling an account — which could include removing, deleting, de-activating or suspending an account — is effectively closing an account which has violated guidelines. This may be temporary or permanent and may be open to redress mechanisms or subject to a specific period of time. It may or may not affect the accessibility of the account's past contributions on the content-sharing service, and may or may not be subject to an obligation to preserve data for law enforcement or similar purposes.
>
> Reporting to law enforcement – A user or account may be reported to a law enforcement agency in order to address illegal activity or imminent risks to safety.
>
> Restricting user privileges – An account may remain operable but with specific privileges restricted, muted, suspended or removed. These privileges may include the ability to live-stream, comment or post.
>
> Warning – A warning message or notice may be issued to an account that has violated company guidelines.

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

> Blocking/disabling – Blocking/disabling means restricting or removing access to specific content for a particular user or group of users. Geo-blocking, for example, restricts access to content for users whose IP addresses are registered within a specific physical location. The content may remain available to some users under specific circumstances.
>
> De-listing – De-listing is the removal by a content-sharing service, or by a user, of content from recommendation lists for users, or from indexing within the "explore" or "discover" functions that allow users to search content on the content-sharing service.
>
> De-monetising – De-monetising content is restricting its ability to leverage the content-sharing service's monetisation features. For example, de-monetising could involve removing the possibility for advertisements to appear alongside content that does not comply with relevant guidelines (e.g. content or others).

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

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

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

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

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

Banning – See Actioning accounts.

Blocking – See Actioning content.

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

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

De-activating – See Actioning accounts.

De-listing – See Actioning content.

De-monetising – See Actioning content.

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

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

1. Proactive – Proactive detection occurs when TVEC or TVEC-related online activity is detected as a result of company-led routine detection. Proactive detection can happen from human, tooling or hybrid systems of review established by a content-sharing service. Proactive detection can be:

    a. Proactive at upload – Proactive detection at upload occurs as soon as a user attempts to add TVEC to, or take specific TVEC-related online actions on a content-sharing service and **before** it is shared with or becomes accessible to others. This is primarily done by automated tools. Once such content or activity is flagged, various moderation actions can take place. For example, if the content is not obviously or overtly against guidelines, it might trigger a triage to human review.

    b. Proactive after upload – Proactive detection after upload occurs after TVEC has been added to a content-sharing service. Depending on the circumstances, this detection may occur **before or after** TVEC has been shared with or become accessible to other users. Again, once TVEC is flagged, various moderation actions can take place.

2. Reactive – Reactive detection occurs when TVEC or TVEC-related online activity is identified through a third-party report made to the content-sharing service. TVEC or TVEC-related online activity may be reported by users (see online community reports below) or by others, such as civil society organisations, governments, law enforcement, trusted notifiers, regulatory bodies, industry bodies, etc. Reports from government institutions or public authorities may take the form of referrals or legal requirements. While there is not always a clear-cut distinction between the two categories, most referrals or legal requirements fall within the parameters contained in the first two items below. Content-sharing services may also have special reporting channels or escalation pathways for specific individuals, entities, types of requests or requirements, TVEC, TVEC-related online activity or situations, such as a real-world terrorist or violent extremist event with direct online implications. The channels or pathways described below may differ or overlap slightly, as they are impacted by how companies design their respective reporting procedures.

    a. Government legal requirements – Government legal requirements direct a content-sharing service to remove TVEC or TVEC-related online activity that violates the law in a national or regional jurisdiction. These requirements may take a number of forms, including notices and orders, and may be founded in various types of laws and legal systems.

    b. Government referrals -- Government referrals are requests by a government institution or public authority to a content-sharing service to review TVEC or TVEC-related online activity on the basis that it may violate the company's community guidelines, terms of service or other relevant guidance documents. The TVEC or TVEC-related online activity may or may not violate local law, as well.

c. Internet Referral Units – Specialised public authorities typically housed within law enforcement bodies, with responsibility for making referrals to content-sharing services. IRUs operate within the confines of their mandate and flag TVEC or TVEC-related online activity that violates a given country's terrorism legislation but which is referred to a company for review against the company's terms of service.

d. Online community reports – Online community reports or flags are a common mechanism for users to report TVEC or TVEC-related online activity to a content-sharing service.

e. Real-world terrorist or violent extremist event with direct online implications – A real-world terrorist or violent extremist event with direct online implications is a concurrent online manifestation of a real-world terrorist or violent extremist incident. It involves TVEC produced by a perpetrator or accomplice that appears to depict ideologically-driven murder (including attempts), torture or serious physical harm and appears to have been designed, produced and disseminated for virality – or has achieved actual virality – being shared online in a manner that presents a threat of unusually high impact (i.e. geographical / cross-platform scale), is likely to cause significant harm to communities, and therefore warrants a rapid, coordinated and decisive response by industry and relevant government agencies. For example, the live-streaming of the Christchurch attack was considered a real-world terrorist or violent extremist event with direct online implications requiring rapid response and action from industry and relevant government agencies.

f. Trusted notifiers – Some content-sharing service designate trusted notifiers or partners who are deemed particularly trustworthy, effective or are subject matter experts in a particular violation or harms type for notifying a content-sharing service of TVEC or TVEC-related online activity that violates its guidelines. Trusted notifier status may include special privileges, for example reports being prioritised, enhanced reporting functionality and increased engagement with the content-sharing service about moderation decisions. Depending on the content-sharing service, trusted notifiers may be comprised of individuals, organisations and/or government institutions.

3. Manual detection – Manual detection (also known as human detection) occurs when people manually identify user-generated TVEC or TVEC-related online activity based on a content-sharing service's guidelines and any relevant internal resources and

processes, including quality control. Depending on the circumstances, these people may be employed, contracted or appointed for this purpose.

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

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

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

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

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

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

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

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

Disabling content – See Actioning content.

Disabling accounts – See Actioning accounts.

Down-ranking – See Actioning content.

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

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

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

Hiding – See Actioning content.

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

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

Moderation – See Detection and moderation, Moderation.

Notification – See Actioning content.

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

Online or digital tooling -- A function, plug-in, or mechanism used to facilitate a particular action or service on a given platform, device or site.

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

Quarantining – See Actioning content.

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

Removing – See Actioning content.

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

Suspending – See Actioning accounts.

Terrorist and violent extremist content (TVEC) – Content refers to any type of digital information, such as text, video, audio and pictures, and the scope of the VTRF is limited to terrorist and violent extremist content (TVEC). As noted in the Origin and Aims section of the VTRF, there is no universally accepted definition of terrorism or violent extremism, nor, by extension, of terrorist and violent extremist content, and no definition is delineated or endorsed here. Instead, that section provides general examples of different approaches that have been taken by different online content-sharing services in defining TVEC, and the metrics ask companies to provide transparency over how they understand the term and any significant updates to that understanding during the reporting period.

Tooling – See Online or digital tooling.

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

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

References

Abdul Rahman Al Jaloud, H. (2019), *Caught in the Net: The Impact of "Extremist" Speech Regulations on Human Rights Content*, https://www.eff.org/wp/caught-net-impact-extremist-speech-regulations-human-rights-content. [4]

Global Research Network on Terrorism and Technology (2019), *Terrorist Definitions and Designations Lists: What Technology Companies Need to Know", Paper No. 7*, https://rusieurope.eu/publication/other-publications/terrorist-definitions-and-designations-lists-what-technology. [5]

Human Rights Watch (2020), *Video Unavailable: Social Media Platforms Remove War Crimes Evidence*, https://www.hrw.org/report/2020/09/10/video-unavailable/social-media-platforms-remove-evidence-war-crimes. [3]

Inter-American Commission on Human Rights (2019), *Resolution 3/2019, Principles on Public Policies on Memory in the Americas*, http://www.oas.org/en/iachr/decisions/pdf/Resolution-3-19-en.pdf. [6]

OECD (2021), "Transparency reporting on terrorist and violent extremist content online : An update on the global top 50 content sharing services", *OECD Digital Economy Papers*, No. 313, OECD Publishing, Paris, https://dx.doi.org/10.1787/8af4ab29-en. [2]

OECD (2020), "Current approaches to terrorist and violent extremist content among the global top 50 online content-sharing services", *OECD Digital Economy Papers*, No. 296, OECD Publishing, Paris, https://dx.doi.org/10.1787/68058b95-en. [1]

United Nations Human Rights Council (2010), *Report of the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism: Ten areas of best practices in countering terrorism*, https://undocs.org/A/HRC/16/51. [7]