March 2025

Basic Online Safety Expectations

Summary of industry responses to mandatory transparency notices addressing terrorist and violent extremist material and activity

Australian Government | eSafetyCommissioner

eSafety.gov.au

Contents

1. Executive summary

On 18 March 2024, the eSafety Commissioner (**eSafety**) gave a non-periodic reporting notice (the **Notice**) to a selection of online service providers: Google, Meta, WhatsApp, Reddit, Telegram and X.[1] The Notice required each to answer questions about the steps it was taking to implement the Basic Online Safety Expectations (the **Expectations**) with respect to **terrorist and violent extremist material and activity** (**TVE**). The Expectations are set by the Australian Government and provided for by the *Online Safety Act 2021* (Cth) (**the Act**).

The Notice was given in accordance with section 56(2) of the Act, which allows eSafety to publish summaries of the information received through notices. We exercise this statutory power in order to improve industry transparency and accountability.

eSafety asked questions about the tools, policies and processes that each of the six companies used to address TVE on their services from 1 April 2023 to 29 February 2024 (the **report period**). In particular, they were required to detail the steps they took to detect and prevent the dissemination of online TVE, mitigate the risks posed by online radicalisation, and safeguard their services from being weaponised to perpetrate and amplify acts of terror and violent extremism.

This transparency report sets out summaries of each service provider's responses to eSafety's questions. It also provides comparisons of the summarised information received about each service, focused on a number of specific issues. These include:

- proactive detection measures
- user reporting
- human moderator resourcing
- efforts to mitigate TVE risks posed by particular service features, such as recommender systems and generative artificial intelligence (AI) capabilities.

A summary of eSafety's key findings from the information provided by industry is available on the eSafety website.

[1] Services covered by the Notices were:
Google – YouTube, Google Drive, Gemini
Meta – Facebook, Messenger, Instagram (including Threads)
WhatsApp – WhatsApp
Reddit – Reddit
Telegram – Telegram
X Corp - X

In addition to these questions, Reddit and Telegram were also asked about the tools, policies and processes they used to detect and address child sexual exploitation and abuse (CSEA)[2] on their respective services. Neither service had previously been required to report on this harm. Google, Meta, WhatsApp and X had previously reported on the steps taken to address child sexual exploitation and abuse, and eSafety published the findings in two transparency reports.

Matter before the Administrative Review Tribunal

X Corp sought review in the Administrative Appeals Tribunal (now the Administrative Review Tribunal) of eSafety's decision to give X Corp the Notice. This matter is ongoing.

Non-compliance with a notice

Telegram failed to provide a response to the Notice given to it by the Notice deadline of 6 May 2024. eSafety subsequently received information from Telegram, five months after the Notice deadline, including some of the information required by the Notice.

As Telegram did not comply with the Notice deadline, eSafety gave it an infringement notice to deter non-compliance in the future. eSafety will continue to use the full range of powers available to it to ensure industry transparency and hold service providers to account.

Significant variation in TVE protection for users

Responses from Google, Reddit, Meta and WhatsApp, as well as information provided by Telegram after the deadline, revealed that although these service providers did have measures in place to detect and address TVE on their services, they were not always applied consistently or comprehensively.

Risks posed by particular service features

Livestreaming and video calling

There is an ongoing risk that TVE can be livestreamed, as happened in the 2019 attack when the murder of multiple people at a Christchurch mosque was broadcast via Facebook to hundreds of users. The online industry made public commitments to prevent TVE livestreaming happening again. The Notice responses revealed the following:

[2] CSEA encompasses both 'child sexual exploitation' (a broad category of content that encompasses material and activity that sexualises and is exploitative to the child, but that does not necessarily involve the child's sexual abuse) and 'child sexual abuse' (which involves sexual assault against a child). Child sexual abuse is a narrower category and can be considered a sub-set of child sexual exploitation.

- Meta had no measures in place to detect livestreamed TVE on Messenger Rooms during the report period. Also, users who were not logged in to Facebook could not report livestreamed TVE to the service using in-service reporting tools.
- WhatsApp did not detect livestreamed TVE during the report period.
- Telegram did not detect livestreamed TVE in Channel livestreams or group video calls.
- Users not logged in to YouTube could not report livestreamed TVE in-service.

Generative AI

There is a risk that generative artificial intelligence (AI) could be misused to create synthetic but highly realistic TVE. The Notice responses revealed the following:

- Google reported it undertook red-teaming (simulation of misuse) on its generative AI service, Gemini, for TVE and for child sexual exploitation and abuse. Despite this, during the report period:
    - Google received 258 user reports about suspected synthetic TVE being generated by Gemini. In the same period, it received 86 user reports of suspected synthetic child sexual exploitation and abuse material generated by Gemini. Google was unable to confirm the number of reports that resulted in confirmation that TVE and child sexual exploitation and abuse material had been generated on Gemini.
- Google also treated TVE differently to child sexual exploitation and abuse material on its Gemini service:
    - Google used **hash-matching tools** on user-uploaded image prompts on Gemini for known child sexual exploitation and abuse material. However, it did not apply the same safety measures for known TVE, despite using TVE hash-matching on YouTube and Drive with hashes sourced from the Global Internet Forum for Countering Terrorism (**GIFCT**[3]).
- Google used classifiers to scan text-based prompts for child sexual exploitation and abuse but not for TVE.

Recommender systems

Without appropriate safeguards recommender systems can support the aim of bad actors who deliberately seek to spread TVE online to glorify the actions of terrorists and violent extremists, promote their hateful ideologies, undermine social cohesion, and jeopardise public safety by inspiring copy-cat attacks[4]. The Notice responses revealed the following:

---

[3] GIFCT, among other things, maintains a database of TVE hashes submitted by member companies, which enable providers to detect when this content is uploaded to their services. https://gifct.org/.

4 eSafety Commissioner, 'Recommender systems and algorithms – position statement', accessed 12 February 2025, URL: https://www.esafety.gov.au/industry/tech-trends-and-challenges/recommender-systems-and-algorithms

- Meta reported that it relied on the removal of TVE from its services in response to questions about interventions in place to prevent the amplification of TVE on Facebook and Instagram. Meta stated 'as TVE is prohibited by the Facebook Community Standards and the Instagram Community Guidelines, our measures are focussed on removing that content [TVE] from our services (rather than preventing its amplification)'.
- Conversely, both Google and Reddit stated that in addition to removing individual items of TVE, they also took proactive measures to limit the recommendation of content that may not be suitable for general audiences.
    - Google said it used teams of human evaluators to train machine learning systems to identify 'borderline content' (which is defined as 'content that comes close to, but does not breach YouTube's Community Guidelines') to limit its amplification by YouTube's recommender system.
    - Reddit said it periodically rated communities based on the content within those communities using an internal taxonomy rating system. Communities must meet a certain size and activity threshold to be eligible for rating, and content from unrated communities is not eligible for recommendation. Content on Reddit also needs to achieve a certain 'suitability score' to be amplified.
- Meta and Google reported staging positive interventions to promote authoritative sources or de-radicalising content on their services.
    - Google said its systems were trained to ensure authoritative sources were elevated in YouTube's search results and recommendations. It also provided 'information panels' on videos and searches 'related to topics that are prone to misinformation'.
    - Meta said that when end-users in Australia searched for words associated with organised hate or violent extremism on Facebook or Instagram, the services promoted a link to resources about how to 'leave violence and extremism behind' as the top search result.
- Telegram was not asked a question about recommender systems but reported that it 'does not employ recommendation algorithms or any other form of targeted amplification'.

End-to-end encryption

When a service is end-to-end encrypted (E2EE) it can limit the automated tools available to detect TVE. The Notice responses revealed the following:

- Meta reported the use of tools to detect and prevent the spread of TVE across parts of its service but not the end-to-end encrypted (E2EE) parts. Meta is in the process of rolling out Messenger to be end-to-end encrypted by default.

- During the report period Meta did not undertake an internal safety risk assessment of its ability to detect and address TVE before implementing E2EE on Messenger and Instagram Direct. Meta did state that it did do broader risk assessment and engagement. (The report period was 1 April 2023 to 29 February 2024, and end-to-end encryption began rolling out for all personal chats and calls on Messenger in December 2023.)
- Meta was reliant on user reports to be able to detect TVE and accounts in breach of its TVE policies on the end-to-end encrypted parts of its service.

User reporting

User reporting options and complaints pathways are important safety measures because they enable users to flag and alert an online service to specific material and activity that is illegal, harmful or otherwise in breach of its terms of service. The time taken to respond to reports of TVE can make a critical difference to its spread and impact. The Notice responses revealed the following:

- The median time service providers took to respond to Australian user reports of TVE varied significantly across services.
    - There was significant variation across Meta's[5] services - taking 0.1 hours to reach an outcome following a TVE report on Messenger (when E2EE enabled or not enabled)[6], 4.2 hours on Facebook Newsfeed, 15.5 hours on Instagram Feed and 59.5 hours on Threads.
    - WhatsApp took 24.13 hours on its E2EE direct messages service, with the only category offered to users to report any high impact or illegal content such as TVE being 'report'.[7]

[5] Meta noted that these figures represent data from 1 October 2023 to 29 February 2024. Meta also reported that the figures were calculated by identifying all user reports on content that was confirmed to violate its TVE policies and 'calculating the 50th percentile of the times taken from the creation of a job to the time an enforcement action was taken'. Meta noted that the creation of a job is when 'a user report cannot be closed automatically (e.g. due to duplication).'

[6] Meta reported that it does not ordinarily track or report data that differentiates when E2EE is and is not enabled regarding response times to user reports that differentiates when E2EE is and is not enabled on Messenger. Meta stated the data provided for this service was 'sourced from non-core datasets and cannot be verified or validated'. It added that 'while Meta has sought to provide accurate data to the best of its ability, Meta has material concerns about the reliability of this data and considers that this data is not sufficiently robust to be used for further analysis'.

[7] WhatsApp reported that these figures reflect enforcement action taken against accounts that were banned for TVE-related violations and had also received a user report over the past 30 days. WhatsApp stated that due to the absence of issue-specific reporting options, WhatsApp cannot identify user reports where the user intended to report TVE specifically. WhatsApp also stated that because it does not log enforcement actions against specific user reports, it was 'not possible … to calculate the median time taken to reach an outcome after receiving a user report of TVE with precision.' WhatsApp reported that these figures are based on the assumption that the 'maximum amount of time' between the user report being made and it being 'enqueued for human review is 24 hours' plus the addition of the time then taken for enforcement action for each service.

- Reddit was also relatively slow to respond, taking 31.3 hours[8].
- Telegram took 18 hours to respond to TVE user reports on Chats and Secret Chats, and 15 hours on Group Chats and Channels.

Proactive detection and blocking

A key principle of eSafety's Safety by Design initiative, and the Expectations, is that safety should be built into a service or feature at the outset, rather than retrofitted after the damage has been done. This is important for the detection and blocking of both **new TVE material** and **known TVE material**. The Notice responses revealed the following:

- WhatsApp rolled out Channels (which is not end-to-end encrypted) during the report period (in June 2023) without implementing hash-matching for known TVE. WhatsApp reported that it only started working on its implementation later in the report period.[9]
- Meta did not use any proactive scanning tools to detect new TVE material on Messenger and Instagram Direct, regardless of whether end-to-end encryption was enabled or not. Notably, in 2022 Meta reported to eSafety that it was using such tools to proactively detect new CSEA material on Messenger and Instagram Direct (when end-to-end encryption was not enabled). Meta was reliant on user reports to detect new TVE on these services.
- Services used tools to proactively detect TVE, though the tools were limited in some cases:
    - Telegram used hash-matching tools on private groups and private channels to detect known TVE, but it did not use tools to detect new TVE on those same parts of the service.
    - Telegram did not use any hash-matching tools on Chats or user reports in relation to Secret Chats.
    - Telegram detected hashes of TVE images and videos it had previously removed from its service, but it did not source hashes of known TVE material from external sources such as GIFCT or Tech Against Terrorism[10].[11]

[8] Reddit noted that users may report material that may be terrorist and/or violent extremist material under the violence reporting option, or potentially under the hate reporting option. Reddit further noted that it has no way to distinguish a user report of TVE from non-TVE violations of these rules, and that it therefore does not have data on the median time taken to reach an outcome after receiving "user reports of TVE" on the service. Reddit also noted that reports that its human safety team determines may relate to terrorist content are sent to a specialised terrorism queue for further human review. The data presented is the median time between a user report and ticket closure for reports escalated to Reddit's specialised terrorism queue.

[9] WhatsApp subsequently advised eSafety that hash-matching tools for TVE on Channels were deployed by May 2024.

[10] Tech Against Terrorism (TAT): A not-for-profit organisation, launched in 2016 by the United Nations. TAT develops technical tools and facilitates knowledge-sharing for countering terrorism and violent extremism online. TAT maintains the Terrorist Content Analytics Platform, accessed 4 June 2024. URL: https://techagainstterrorism.org/terrorist-content-analytics-platform

[11] Following consultation with Telegram on the proposed report for publication, Telegram reported that it 'routinely reviewed hash databases compiled by Europol to inform its systems for proactive detection.'

- Google only used hash-matching to detect exact matches of TVE content, rather than edited copies. This is concerning in the context of the number of variations of the Christchurch video - Facebook reported 800 different versions in the first days after the attack.[12]

- Service providers were broadly blocking 'join-links'[13] and URLs linking to websites dedicated to TVE and to known TVE on other websites, with some exceptions.
    - WhatsApp, which is end-to-end encrypted, and Meta's end-to-end encrypted services did not block them.
    - Telegram did not block 'join-links' and URLs to TVE across any parts of its service.
    - While Meta did not block URLs linking to known TVE on end-to-end encrypted parts of its service, it did use an on-device functionality called 'Safe browsing' that detects URL snippets in its end-to-end encrypted messaging services. Users are warned about potential issues with the links.
    - While Google did block 'join-links' and URLs on YouTube, it did not source URLs for known TVE from external sources. eSafety notes that Google is a member of GIFCT, and although it took hashes of known TVE material from the GIFCT database, it did not source URLs to known TVE from GIFCT.

Trust and safety staff and language coverage

The number of trust and safety workers, along with the language skill set of moderators, can impact the ability to address TVE.

Staffing levels

The Notice responses revealed the following:

- There was a 27.8% reduction in Meta trust and safety staff employed (other than engineers and content moderators) from 31 March 2023 to 31 December 2023. The number of content moderators contracted by Meta fell by 10.6% over the same period.
- There was a 10.7% reduction in Google trust and safety staff employed (other than engineers and content moderators). The number of content moderators employed by Google increased by 7.9%.

[12] Meta, 'A further update on New Zealand terrorist attack', 2019, accessed 10 October 2024, URL: https://about.fb.com/news/2019/03/technical-update-on-new-zealand
[13] A feature on some messaging services that enables end-users to forward and share access to private groups.

Language coverage

There were significant differences between services in terms of the languages covered by human moderators and **automated tools**:

- Reddit and WhatsApp human moderators only covered 13 and 6 languages respectively and only 1 of the top 5 languages, other than English, spoken in Australian homes, despite their high use in Australia[14][15]. In contrast, Google covered approximately 80 languages and Meta 109 languages, including all top 5 languages, other than English, spoken in Australian homes (Arabic, Cantonese, Mandarin, Vietnamese and Punjabi).[16] Telegram covered 47 languages, but only 2 of the top 5 languages, other than English, spoken in Australian homes.[17]
- The tools used by Google (YouTube) and Meta to detect phrases, codes and hashtags indicating likely TVE in text operated in upwards of 100 languages whereas Reddit's tools operated in 27 languages across some parts of its service.

Volunteer moderation

- Meta (Facebook), Reddit and Telegram used volunteer moderators to enforce service-wide policies as well as community-specific rules with a range of moderation tools. However, trust and safety staff were not automatically informed when volunteer moderators removed an account for a TVE violation.

Recidivism

Banned or suspended users who use new details to re-register with an online service, or register with an alternative one, can continue to be a TVE risk. The Notice responses revealed the following:

- The extent of measures to address **recidivism** varied considerably:
    - Google's Drive, Telegram, and WhatsApp had minimal measures in place to detect recidivism of users and groups, channels or communities.

[14] Digital 2023 Australia (February 2023), Jan 2023 most used social media platforms, accessed 6 August 2024, URL: https://www.slideshare.net/slideshow/digital-2023-australia-february-2023-v01/255754526?from_search=0#57

[15] WhatsApp subsequently stated: 'In addition, WhatsApp provides its reviewers with translation tools to enable them to review material in languages other than their native languages'.

[16] Australian Bureau of Statistics, 'Cultural diversity: Census', 28 June 2021, URL: https://www.abs.gov.au/statistics/people/people-and-communities/cultural-diversity-census/latest-release#:~:text=Top%205%20languages%20used%20at,Punjabi%20(0.9%20per%20cent).

[17] Telegram reported that since the report period, it had expanded the languages covered by its contracted content moderators by adding Afrikaans, Bengali (Bangladesh), Chichewa (Zambia), Dhivehi (Maldives), Dutch, Gujarati, Kabyle (Algeria), Kinyarwanda, Lithuanian, Macedonian, Sinhalese (Sri Lanka), Thai and Punjabi.

- Instagram and Facebook mutually shared information about accounts banned for TVE and also shared information with WhatsApp for 'severe violations of our DOI [Dangerous Organisations and Individuals[18]] and other relevant policies'.
- Conversely, WhatsApp did not share any information with Facebook or Instagram about accounts banned for TVE.

Meta's Dangerous Organisations and Individuals List

- WhatsApp did not prohibit all organisations on Meta's Dangerous Organisations and Individuals List from using WhatsApp's private messaging service.

Account bans

- eSafety considers that Google's approach on Drive to limit account bans to accounts that are 'owned or operated by a known terrorist or violent extremist organisation' may result in terrorists and violent extremists who are not associated with a specific organisation – such as the Christchurch attacker – evading a ban.[19]

Information received about child sexual exploitation and abuse

Reddit and Telegram were also asked questions about measures to detect and address child sexual exploitation and abuse material and activity on their services.

User reporting

- It took Reddit almost double the time to action an Australian user report of child sexual exploitation and abuse on public subreddits (12.4 hours)[20] compared to private subreddits (6.8 hours)[21]. Reddit took more than 24 hours to respond to Australian user reports of child sexual exploitation and abuse in Channels (29.5 hours)[22].

[18] Meta, 'Dangerous organisations and individuals', URL supplied by Meta on 24 June 2024, URL: https://transparency.meta.com/en-gb/policies/community-standards/dangerous-individuals-organizations/
[19] The Christchurch attack led to a system, set up by the GIFCT and of which Google is a member, for dealing with material that is not associated with a specific terrorist group.
[20] Australian data.
[21] Australian data.
[22] Australian data.

- It took Telegram 11 hours to action user reports about child sexual exploitation and abuse in Chats and Secret Chats, and 10 hours in Channels and Group Chats (irrespective of whether they were public or private).[23]

Proactive detection and blocking

- There was inconsistent use of tools for detecting known and new child sexual exploitation and abuse:
    - Reddit used **hash-matching tools** to detect known child sexual exploitation and abuse images and videos across all parts of the service asked about in the Notice.
    - Although Reddit used tools to detect new child sexual exploitation and abuse images and videos, these tools did so based on the text included in the image, video and video post (such as the post title) and not through other indicators in the image or video (such as nudity detection and age estimation). This may mean key indicators of child sexual exploitation and abuse material were missed.
    - Telegram used tools to detect known and new child sexual exploitation and abuse images and videos, but not consistently across the service. It used hash-matching tools to detect known images and videos on private groups and private channels (which are not end-to-end encrypted), but did not use any tools to detect new images and videos on those same parts of the service. Telegram did not use tools to proactively detect known or new child sexual exploitation and abuse images or videos on Chats or in user reports about content in Secret Chats (neither of which is end-to-end encrypted).
    - Telegram detected hashes of child sexual exploitation and abuse images and videos it had previously removed from its service, but it did not source hashes of known images and videos from external sources such as NCMEC or the Internet Watch Foundation.[24]
    - Reddit used language analysis tools to detect terms, abbreviations, codes and hashtags indicating likely child sexual exploitation and abuse activity, such as grooming, sexual extortion and the trading and sale of child sexual exploitation and abuse material, on most but not all parts of its service.
    - Telegram was also inconsistent in its use of language analysis tools to detect child sexual exploitation and abuse activity. Tools were used on some parts of the

[23] Telegram stated that it calculated these figures the net time frames between the submission of each report and the moderator's decision with respect to that report. Telegram also stated that it did not have the capability to provide Australia-specific data.
[24] Telegram since reported that, as at October 2024, it was 'in the process of joining the Internet Watch Foundation's safety programs involving, inter alia, access to URL lists containing links to known CSAM websites'.

service, but not Chats, user reports in Secret Chats, or private channels and group chats.

- The service providers took different approaches to the blocking of links to known child sexual exploitation and abuse material:
    - Reddit blocked URLs.
    - Telegram did not block URLs.
- There were also differences in the approach to detect recidivism related to child sexual exploitation and abuse:
    - Reddit used multiple indicators.
    - Telegram used a minimal number of indicators.
- For Reddit, there was considerable variation in detection of child sexual exploitation and abuse by proactive tools compared to material reported by users, trusted flaggers and others across Reddit's services – even though the same automated tools were used on both Chat and Channels and the same reporting categories were offered to users.
    - More than 90% was detected by proactive tools on Chat.
    - More than 80% was reported by users, trusted flaggers or others on Channels.

Furthering transparency

eSafety hopes that the information in this report (and other transparency reports) will be used by the services named and all other industry participants to address key online safety challenges, and encourage greater transparency in the future, particularly regarding TVE (and child sexual exploitation and abuse).

For its part, eSafety will:

- use the information gathered from the Notice to continue to build an understanding of industry practices, with a focus on improving transparency and accountability
- raise specific gaps and vulnerabilities with service providers that received the Notice, to understand more about why certain aspects of the Expectations may not currently be complied with and any future steps that are planned to ensure their services are implementing the Expectations.

2. Glossary

The following glossary of terms has been provided for the benefit of the reader of this report.

- **Audio and/or video classifiers:** Artificial intelligence used to sort information into categories.
- **Automated tools:** Technology used to detect harmful or illegal material and activity. In the context of this report, these tools are used to support content moderation actions and decisions.
- **Expectations:** The expectations set out in the Online Safety (Basic Online Safety Expectations) Determination 2022.[25]
- **GIFCT:** The Global Internet Forum for Countering Terrorism.[26]
- **Hash-matching tools:** Digital technology that is used to create a hash of an image or video which is then compared against hashes of other photos to find copies of the same image or video.
- **Known TVE material:** Images/videos/written material that have previously been confirmed to contain TVE, such as that captured in the GIFCT hash-sharing database.
- **Livestreamed TVE:** Transmission or receipt of TVE material or activity live via webcam or video to people anywhere in the world. Livestreaming includes one-on-one video calls and video calls where one or multiple people stream material to a group of any size.
- **New TVE material:** New TVE images are images that have not been previously confirmed, hashed, and stored in a hash database.
- **Notice:** Non-periodic reporting notice given to an online service provider under section 56(2) of the Act on 18 March 2024.
- **Purple/violet-teaming:** A collaborative approach to penetration testing where adversarial (red team) and defensive (blue team) teams work together to probe, refine, and strengthen defences against realistic simulated attacks.
- **Recidivism:** Banned or suspended users re-registering to an online service with new details to continue perpetrating online abuse.

[25] On 30 May 2024, The Minister for Communications amended the Expectations to address changing online safety challenges by strengthening the Expectations and articulating additional reasonable steps that providers can take to meet them. The findings in this transparency report reflect information relating to the period 1 April 2023 to 29 February 2024. This period preceded the amendments to the Expectations.

[26] GIFCT, 'Preventing terrorists and violent extremists from exploiting digital platforms', accessed 18 December 2024, URL: https://gifct.org/.

- **Recommender algorithms:** The set of computing instructions that determine what a user will be served based on many factors. This is done by applying machine learning techniques to the data held by online services, to identify user attributes and patterns and make recommendations to achieve particular goals.
- **Report period:** When online service providers receive a Notice from eSafety they are required to prepare a report about the extent to which they complied with the Basic Online Safety Expectations during a specified period. This period is referred to as the report period. The report period for this set of Notices is 1 April 2023 to 29 February 2024. Information provided should reflect this period, unless stated otherwise.
- **Terrorist and violent extremist material and activity (TVE):** Unless otherwise specified, 'TVE' refers to terrorist and violent extremist material and activity[27]. (Some questions to services providers asked about material only, or specific kinds of material, such as images. Service providers were asked to respond to questions in relation to TVE using their closest equivalent definitions in their terms of service, guidelines and policies.)
- **The Act:** The *Online Safety Act 2021* (Cth).
- **The Determination:** The *Online Safety (Basic Online Safety Determination) 2022* (Cth).[28]
- **Trusted flagger:** An individual or entity which is considered to have particular expertise and responsibilities for the purposes of tackling harmful content online.

[27] This may include but is not limited to material or activity that:
   a. depicts or includes a 'terrorist act' as defined in section 100.1 of the Criminal Code (Cth) no matter where the action occurs, the threat of action is made, or where the action would occur if carried out;
   b. depicts or includes advocating the doing of a 'terrorist act', e.g. 'pro-terror material', as defined in the Consolidated Industry Codes of Practice for the Online Industry (Class 1A and Class 1B Material) Head Terms – Annexure A ;
   c. depicts or includes promoting, inciting or instructing in matters of crime or violence with the intention of advancing a political, religious or ideological cause;
   d. has the effect of – whether intentionally or unintentionally – promoting or glorifying material or activity that is underpinned by violent extremist or terrorist ideologies; or
   e. promotes or celebrates terrorist leaders, organisations and groups, their actions or ideologies.

Not all material or activity that falls within these, or other, categories will constitute TVE. For example, see the defences that apply to the access of abhorrent violent material at section 474.37 of the Criminal Code, which includes defences for news reports, and scientific, medical, academic or historical research, amongst others.

[28] Amended by the Minister on 30 May 2024, after the conclusion of the report period.

3. Information about the Notice

The Basic Online Safety Expectations

The Basic Online Safety Expectations Determination 2022 sets out the Australian Government's Expectations that social media, messaging, gaming, dating, file sharing services and other apps and websites will take reasonable steps to keep Australians safe online.

Compliance with the Expectations is not enforceable, but eSafety can require service providers to report on the steps they are taking to meet the Expectations. There are financial penalties for service providers that do not comply with a Notice.

Further information on the Expectations and associated powers can be found in eSafety's Regulatory Guidance.

The Expectations work alongside Australia's online industry Phase 1 Codes and Standards which place mandatory and enforceable obligations on relevant participants in the online industry requiring them to take action to reduce access and exposure to illegal content, including some forms of TVE.

Who received the Notice?

eSafety gave notices to the following six service providers, under section 56(2) of the Act:

Provider that received the section 56(2) Notice	Services
Google LLC.	YouTube Google Drive Gemini
Meta Platforms, Inc.	Facebook Messenger Instagram (including Threads)
WhatsApp LLC.	WhatsApp
Reddit Inc.	Reddit
Telegram FZ LLC.	Telegram
X Corp.	X

In deciding which service providers will receive a Notice, eSafety is required to consider several criteria specified in the Act:

- the number of complaints eSafety has received under the Act in relation to the service in the previous 12 months

- any previous contraventions of civil penalty provisions related to reporting on the Expectations
- any deficiencies in the service provider's safety practices and/or terms of use
- whether the service provider has agreed to give the Secretary regular reports relating to safe use of their service
- any other matters the Commissioner considers relevant.

Examples of other matters that eSafety has said in the Basic Online Safety Expectations Regulatory Guidance it may take into account include:

- a service's reach and the profile of its users, including whether it is used by children
- the measures the service provider currently has in place to protect users from harm
- the information already published by a service provider and any absence of information regarding a service's safety policies, processes and tools, or limited information about the impact or effectiveness of these interventions
- aggregated evidence from eSafety's other regulatory schemes, such as types of complaints, a service provider's responsiveness to removal requests/notices, or other investigative insights regarding service safety issues
- evidence of systemic harm, or evidence of key safety risks, relative to the Expectations, including from victims, charities, media, academics or other experts.

The choice of service providers that receive notices is not, in itself, indicative of eSafety's views or level of concern with those service providers' compliance with the Expectations. There may be service providers with material accessible in Australia that are more, or less, compliant with the Expectations than the service providers who received Notices.

What questions did eSafety ask?

The Notice required service providers to respond to eSafety in the manner and form specified in the Notice. This involved responding to a set of specific questions, using a template provided by eSafety. The questions were a mix of yes and no questions, and questions allowing free text answers or seeking specific data. eSafety's view is that targeted questions assist both the service provider and eSafety and ensure the provision of meaningful information.

Through answering the questions, providers were required to report on the specific steps they were taking to meet the relevant Expectations by detecting and preventing TVE (and in some cases, child sexual exploitation and abuse material and activity) on their services.

Service providers were not asked the same questions in every instance. Each Notice required the service provider to respond to a unique set of questions tailored to the specifics of their services, the relevant risks and any information gaps about their safety practices.

An overview of the types of questions eSafety asked is contained in the following table, with the corresponding Expectation(s) listed:

Areas covered by Notices	Corresponding Expectation in Determination
The definitions that service providers use to describe and categorise 'terrorist' and 'violent extremist' material and activity for purposes of content moderation.	Section 14 (Providing terms of use and certain policies and procedures regarding reports, complaints and conduct)
The extent to which service providers use automated tools to proactively detect TVE on their services, including known 'hashed' TVE material, new 'first-generation' TVE material, and livestreamed TVE.	Section 6(2) (Ensuring reasonable steps to proactively minimise the extent to which material or activity on the service is unlawful or harmful) Section 11 (Minimising provision of certain material)
The extent to which service providers are detecting and addressing TVE on encrypted services.	Section 8 (If the service uses encryption, the provider of the service will take reasonable steps to develop and implement processes to detect and address material or activity on the service that is unlawful or harmful)
Availability of mechanisms for users to report TVE on the services.	Section 13 (Providing mechanisms to report and make complaints about certain material (including forms of TVE material)) Section 14(1)(c) (Policies and procedures for dealing with reports and complaints mentioned in section 13 or 15) Section 15 (Providing mechanisms to report and make complaints about breaches of terms of use)
The resources service providers deploy to support content moderation by humans on their services, including the resourcing of expertise in TVE issues.	Section 6 (Ensuring safe use) Section 11 (Minimising provision of certain material)
Steps taken to prevent banned or suspended users from creating new accounts (recidivism).	Section 6(2) (Ensuring reasonable steps to proactively minimise the extent to which material or activity on the service is unlawful or harmful) Section 9 (Preventing anonymous accounts being used for unlawful or harmful material or activity) Section 11 (Minimising provision of certain material)

	Section 14(2) (Reasonable steps to ensure that penalties for breaches of its terms of use are enforced against all accounts held or created by the end-user who breached the terms of use of the service)
Steps taken to avoid the risk of amplifying harmful content through recommender systems.	Section 6 (Ensuring safe use) Section 11 (Minimising provision of certain material)
Steps taken to mitigate the risk of generative AI being misused to perpetrate harm.	Section 6 (Ensuring safe use) Section 11 (Minimising provision of certain material) Section 13 (Providing mechanisms to report and make complaints about certain material (including forms of TVE & CSEA material)) Section 14(1)(c) (Policies and procedures for dealing with reports and complaints mentioned in section 13 or 15) Section 15 (Providing mechanisms to report and make complaints about breaches of terms of use)
The extent to which service providers use automated tools to proactively detect CSEA on their services (for Telegram and Reddit).	Section 6 (Ensuring safe use) Section 11 (Minimising provision of certain material)

What was the Notice process?

Service providers had 49 days to respond, or longer as agreed with eSafety. Several extensions were granted where requested by service providers. Service providers were invited to discuss with eSafety any questions they had about the Notice, how to respond, or the scope of the questions.

What process was followed once the information was received?

Assessment and follow-up questions

On receipt of service provider responses, eSafety assessed if each service provider had answered the questions required by the Notice.

Where the service provider's response was not clear, eSafety followed up to seek clarification of the response and any further information the service provider opted to supply to give context.

Service providers were invited to discuss with eSafety any questions they might have.

Draft summary reports

Service providers were given a draft of their individual summary report and summary tables relating to their service(s) prior to publication. Service providers were invited to discuss with eSafety the proposed publication, any concerns they might have, and any submissions they wished to make about information included in the summaries. eSafety considered all submissions received from service providers to finalise this transparency report.

What information has been published, and what has been excluded?

This report summarises the information that eSafety received from responses to the Notice by Google, Meta, Reddit and WhatsApp (while WhatsApp is owned by Meta, it is considered a separate service provider for the purposes of the Basic Online Safety Expectations, so it was given a separate Notice). It also includes some information provided by Telegram after the deadline.

In addition, the report sets out comparisons of the summarised information received about each service, focusing on a number of specific issues.

The summaries in this report do not reflect service providers' responses in their entirety. In line with eSafety's regulatory guidance, certain information has been withheld where eSafety considered it was not appropriate to disclose – for example, because it contained commercial-in-confidence information or because publication of the information would not serve the public interest.

In particular, eSafety has determined that it is not in the public interest to publish specific indicators and signals that service providers deploy to detect users seeking to commit crimes and cause harm, and to prevent recidivism. eSafety engaged with law enforcement agencies and other counter-extremism and child safety experts to seek views on what kind of information would not be in the public interest to publish.

A summary of eSafety's key findings from the information provided by industry is available on the eSafety website.

The following points should also be noted:

- The information provided in responses to the Notice has not been verified by eSafety, although service providers are required to respond truthfully and accurately. Information is published in the interests of transparency and accountability.

- The information summarised in this report is based on the responses eSafety received, which reflect a particular period in time – the period 1 April 2023 to 29 February 2024 inclusive, or other periods within this timeframe as specified. Service providers may have implemented changes to tools, policies and processes since this information was provided.
- All data is global, unless otherwise stated.
- Bolded terms are defined in the glossary of this report, unless otherwise stated.

Matter before the Administrative Review Tribunal

X Corp sought review in the Administrative Appeals Tribunal (now the Administrative Review Tribunal) of eSafety's decision to give X Corp the Notice. This matter is ongoing.

What happens next?

The information presented in this summary provides new insight into the steps that these service providers are taking to address online TVE. eSafety hopes that the information obtained from this Notice (and other transparency notices) will be used by the services named and all other industry participants to address key online safety challenges, and encourage greater transparency in the future – including through their own voluntary disclosures.

eSafety's Basic Online Safety Expectations Regulatory Guidance sets out our planned approach to the exercise of our powers in respect of the Expectations more generally. In the coming months eSafety will:

- use the information gathered from the responses to the Notice to continue to build an understanding of industry practices, with a focus on improving transparency and accountability around online TVE (and child sexual exploitation and abuse material and activity)
- raise specific gaps and vulnerabilities with service providers that received the Notice to understand more about why certain aspects of the Expectations may not currently be complied with, and any future steps that are planned to ensure their services are implementing the Expectations, particularly regarding TVE (and child sexual exploitation and abuse material and activity)
- continue to engage with service providers who received the first periodic reporting notices (in July 2024) focussed on acute harms and potential deficiencies in their safety processes – the intent of periodic reporting notices is to track key safety issues and progress against them

- continue to engage with service providers more generally in relation to the Expectations, including through the first round of information requests given in September 2024 under section 20 of the Determination
- continue to expand use of non-periodic notices to other acute harms – we welcome input from all stakeholders on the areas where greater transparency is needed.

eSafety also intends that the information in this report, together with other transparency reports, will be used by researchers, academics, the media and the public to scrutinise the efforts of industry, in order to improve accountability and encourage implementation of the Expectations.

4. Compliance with the Notice and action taken by eSafety

eSafety's powers to require reports

Information is sought through non-periodic reporting notices to improve transparency and accountability, incentivise improvements in safety standards, and help eSafety to determine whether a service provider is compliant with the Expectations.

A non-periodic reporting notice requires the service provider to prepare a report about compliance with one or more of the Expectations, prepare the report in the manner and form set out in the notice, and to provide it to eSafety[29]. Service providers are required to comply with a notice to the extent they are capable[30].

When this Notice was given in March 2024, eSafety also supplied each service provider with a response template with questions tailored to that provider and its services, related to specific Expectations. The Notice required responses to all these questions.

Service providers were required to respond by the deadline set by the Notice. eSafety informed each service provider that they could request an extension of time to enable them to comply with the Notice. eSafety also informed each service provider that it should contact eSafety if it had any questions about the Notice, the information being sought, or how to respond.

In addition, service providers were notified that they had the right to seek an internal or external review of the decision to give them a notice under section 56(2) of the Act. Information on the different review options available was included with each Notice. Internal review is a

[29] Section 56(2) of the Act.
[30] Section 57 of the Act.

review conducted under eSafety's Internal Review Scheme. At the time the Notice was given, external review would be a review conducted by the Administrative Appeals Tribunal (later superseded by the Administrative Review Tribunal), as well as other options.

Why it is important that service providers comply with transparency notices

Service providers are required to comply with their legal obligations under Australian law, including the Online Safety Act.

A service provider's failure to comply with a reporting notice deadline prevents eSafety from obtaining information about the steps it is taking to comply with the Expectations, as intended by the Act. This limits the transparency of service providers, prevents them from being held accountable and impacts eSafety's ability to effectively fulfil its statutory functions in a timely manner.

This Notice was related to serious and egregious harms – TVE and, in the case of Reddit and Telegram, child sexual exploitation and abuse – and all Australians have a right to know how service providers are protecting the safety of users and the general public.

It is service providers themselves who hold the information about the internal tools, policies and processes they use to detect and address these harms. The Act recognises this and provides eSafety with powers to mandate the provision of information. The importance of these powers is recognised by the provision of civil penalties where a service provider fails to comply to the extent they are capable. Parliament did not intend for these powers to be voluntary requests for information.

The transparency and accountability objectives of the Act aim to promote the online safety of Australians by increasing awareness of online safety issues and the way that services respond to online harms. These objectives incentivise improvements and encourage best practice in the safety measures taken by industry.

In order for the objectives of the Act to be met, it is important that service providers comply with statutory notices by the deadline, provide complete and accurate information, and are deterred from non-compliance.

Finding of non-compliance

eSafety considers that Telegram did not comply with the Notice given to it for the following reason:

- Telegram did not provide a response to the Notice by the deadline of 6 May 2024.

Telegram did not engage with eSafety during the Notice period to seek any clarification that might have enabled compliance.

Telegram's non-compliance with the Notice deadline delayed public transparency and accountability, and obstructed eSafety from delivering its functions under the Act.

eSafety advised Telegram that it had failed to respond to the Notice and gave it further opportunity to provide the information, or reasons why the information could not be supplied. eSafety subsequently received information from Telegram that was required by the Notice, five months after the deadline, on 13 October 2024.

eSafety has given Telegram an infringement notice of $957,780. Telegram has 28 days to request the withdrawal of the infringement notice or to pay the penalty. If Telegram chooses not to pay the infringement notice, it is open to the Commissioner to take other action.

5. Transparency: Responses by issue

Service providers were required to report on the measures they were taking during the report period to address various types of TVE on their services.

Where services providers reported on information that addressed the same or similar issues, eSafety has compiled that information in summary tables. Setting it out in this way allows easy comparison, which enables a fuller understanding of the differences in how the services operated. The information reflects a point in time, and eSafety acknowledges that the tools, policies and processes may have since changed and may continue to change.

eSafety also recognises that each provider and service is different – with different functionality, architectures, business models and user bases. This means an intervention or tool which may be proportionate and appropriate on one service, may not be on another. When reviewing the tables it is important to take into account the nature of the service and the context in which the service operates, as well as the risk of online harms associated with that service.

In this section eSafety also explains why it asked questions related to particular issues and gives a high-level overview of the technologies available to industry.

eSafety intends that this report is a useful transparency and accountability tool that provides information about the actions service providers are taking to keep all Australians safe online. These tables do not reach a conclusion about the appropriateness of actions taken by providers, or a conclusion regarding their compliance with the Expectations.

Defining 'terrorist' and 'violent extremist' material and activity

eSafety recognises that there is no universally accepted definition of 'terrorism' or 'violent extremism', nor of terrorist and violent extremist material (or content) and activity (or conduct). 'TVE' is an abbreviation commonly used by the online industry and related stakeholders to refer to both the material and activity, so it is used in this report.

eSafety asked service providers to report on safety measures taken during the report period to protect Australians from online TVE and the risk of harm that such material and activity poses to the safety and security of Australians. To help guide and align the framing of each service provider's response to the Notice, eSafety gave the following context to consider when answering the Notice questions:

> '[TVE] may include but is not limited to material or activity that:
>
> a) depicts or includes a 'terrorist act' as defined in section 100.1 of the Criminal Code Act 1995 (Cth) no matter where the action occurs, the threat of action is made, or where the action would occur if carried out;
>
> b) depicts or includes advocating the doing of a 'terrorist act', e.g. 'pro-terror material', as defined in the Consolidated Industry Codes of Practice for the Online Industry (Class 1A and Class 1B Material) Head Terms – Annexure A;
>
> c) depicts or includes promoting, inciting or instructing in matters of crime or violence with the intention of advancing a political, religious or ideological cause;
>
> d) has the effect of – whether intentionally or unintentionally – promoting or glorifying material or activity that is underpinned by violent extremist or terrorist ideologies; or
>
> e) promotes or celebrates terrorist leaders, organisations and groups, their actions or ideologies.
>
> Not all material or activity that falls within these, or other, categories will constitute TVE. For example, see the defences that apply to the access of abhorrent violent material at section 474.37 of the Criminal Code, which includes defences for news reports, and scientific, medical, academic or historical research, amongst others.'

In addition to providing this contextual framing, eSafety asked each service provider to explain how they defined the terms 'terrorist material and activity' and 'violent extremist material and activity' for the purposes of their own terms of service and community guidelines.

These questions were asked to acquire a deeper level of insight and understanding of how service providers interpreted these terms and concepts for each of their services and applied the interpretations in the content moderation policies and practices on each.

Each service provider's definition of 'terrorism' or 'violent extremism' impacts the decisions made about the kinds of content, conduct, and entities included or excluded from the scope of permissible material and activity on their respective services. The drawing of these boundaries affects the decisions that content moderators, and other trust and safety personnel, make when they consider enforcement action against material or activity that potentially meets the standard of harmful TVE. These decisions have direct implications for the safety of users on the service, as well as broader implications concerning the moderation of speech online.

Details of how service providers defined 'terrorist' and 'violent extremist activity' for the purposes of their terms of service, community guidelines or other equivalent service rules can be found in individual service provider summaries in section 6.

Proactive detection

Proactive detection encompasses a broad range of interventions that service providers may take to discover and take action against material or activity on a service before it is reported by a user. These interventions typically involve the use of technologies and tools to automatically scan for material or activity that is prohibited by a service's terms of service.

Detecting known material using hash-matching tools

Service providers were asked about their use of hash-matching to detect various forms of 'known' TVE material. Known TVE material is material that has been previously assessed and verified as TVE material. Hash-matching tools work by creating a unique digital signature (known as a 'hash') of an image or video which is then compared against signatures ('hashes') of other photos or videos to find copies of the same material. Hash-matching allows online service providers to detect and remove images or videos containing unlawful or seriously harmful material – such as TVE material – without needing to store and refer to original copies of the material itself.

Service providers may maintain their own internal databases of TVE hashes, or they may submit and receive hashes from organisations that specialise in collating hashes of material detected by other service providers. For example, GIFCT operates the Hash Sharing Consortium which

provides its members with a database of hashes of images, videos, PDFs and URLs known to contain TVE material.

Hash-matching enables service providers to prevent the re-upload of copies of known TVE material at scale and with a high degree of accuracy. For example, PhotoDNA, an image-hashing tool developed by Microsoft and Dartmouth College in 2009, has a reported error rate of 1 in 50 billion.

There is a broad range of hash-matching tools available to the online industry. PhotoDNA and Facebook's TMK+PDQ are examples of existing tools, made available to organisations. Previous transparency reports published by eSafety have also revealed that some companies have developed their own tools for detecting known child sexual exploitation and abuse material.

eSafety asked about the detection of known TVE material, in relation to sections 6(2) and 11 of the Determination.

Table 1: In response to the notices, the following information was given by service providers regarding the use of hash-matching tools to identify images containing known TVE material.

Provider	Services/parts of services	Used image hash- matching tools	Names of tools used
Google	**YouTube** • YouTube • YouTube profile pictures • YouTube video thumbnails	Yes	MD5/SHA256
	Drive (consumer version; stored content)	No	
	Drive (consumer version; content when it is shared)	Yes	MD5/SHA256
Meta	**Facebook** • Facebook Newsfeed posts, including comment sections • Facebook Group (public) posts, including comment sections • Facebook Group (closed/private) posts, including comment sections • Facebook Channels • Facebook Stories • Facebook profile pictures • Facebook Group profile pictures	Yes	• SimSearchNet++ • PhotoDNA • PDQ
	Messenger • Messenger (when E2EE not enabled) • Messenger Group cover photos	Yes	• PhotoDNA • PDQ

	• Messenger Channels • Messenger Stories		
	Instagram • Instagram Feed • Instagram Direct (when E2EE not enabled) • Instagram profile pictures • Instagram Groups • Instagram Groups profile pictures • Instagram Reels	Yes	• SimSearchNet+ • PhotoDNA • PDQ
	Messenger • Messenger (when E2EE enabled) **Instagram** • Instagram Direct (when E2EE enabled)	No	
	Threads • Threads • Threads profile picture	Yes	• SimSearchNet+ • PhotoDNA • PDQ
Reddit	• Subreddits (public) • Subreddits (private)	Yes	• Snooron – Internal hash-matching functionality • Rule-Executor-V2 (REV2) – automated enforcement system
	• Chat • Channels	No (but since implemented)	Implemented since reporting period: • Snooron– Internal image hash-matching functionality • Rule-Executor-V2 (REV2) – automated enforcement system
	• Channel profile picture • Account profile picture • Subreddit profile picture	No	Reddit stated it is 'currently building new internal hash tooling which will supplement detection' in these parts of its service.
WhatsApp	• Content in user reports • User profile picture • Groups profile picture • Communities profile picture	Yes	Media Match Service

	• Channels messages	No (but since implemented)	
	• Channels profile picture • Status	No	
Telegram	• Group chats (public) • Group chats (private) • Channels (public) • Channels (private) • Stories • User profile picture • Group profile picture • Channel profile picture • Content in user reports	Yes	Internal Telegram Hash Matching System
	• Chats • Secret chats (user reports)	No	

Table 2: In response to the notices, the following information was given by providers regarding the use of hash-matching tools to identify videos containing known TVE material.

Provider	Services/parts of services	Used video hash-matching tools	Names of tools used
Google	**YouTube** • YouTube	Yes	• MD5/SHA256
	Drive (consumer version; stored content)	No	
	Drive (consumer version; content when it is shared)	Yes	• MD5/SHA256
Meta	**Facebook** • Facebook Newsfeed posts, including comment sections • Facebook Group (public) posts, including comment sections • Facebook Group (closed/private) posts, including comment sections • Facebook Channels • Facebook Stories	Yes	• Proprietary Meta video hashing tool • VideoMD5
	Messenger • Messenger (when E2EE not enabled) • Messenger Channels • Messenger Stories	Yes	• Proprietary Meta video hashing tool
	Messenger • Messenger (when E2EE enabled)	No	

	Instagram • Instagram Direct (when E2EE not enabled) • Instagram Groups	Yes	• Proprietary Meta video hashing tool • VideoMD5
	Instagram • Instagram Feed • Instagram Reels	Yes	• Proprietary Meta video hashing tool • VideoMD5 • VideoPDQ
	Instagram • Instagram Direct (when E2EE enabled)	No	
	Threads • Threads	Yes	• Proprietary Meta video hashing tool • VideoMD5 • VideoPDQ
Reddit	• Subreddits (public) • Subreddits (private)	Yes	• Snooron – Internal hash-matching functionality • Rule-Executor-V2 (REV2) – automated enforcement system
WhatsApp	• Content in user reports	Yes	• Media Match Service
	• Channel messages	No (but since implemented)	
	• Status	No	
Telegram	• Group chats (public) • Group chats (private) • Channels (public) • Channels (private) • Stories • Content in user reports	Yes	Internal Telegram Hash Matching System
	• Chats • Secret chats (user reports)	No	

Table 3: In response to the notices, the following information was given by providers regarding the use of hash-matching tools to identify written material (for example manifestos or text promoting, inciting, instructing terrorism) containing known TVE material.

Provider	Services/parts of services	Used hash-matching tools to identify known written material	Names of tools used
Google	**Drive** (consumer version; stored content)	No	
	Drive (consumer version; content when it is shared)	Yes	• MD5/SHA256
Meta	**Facebook** • Facebook Newsfeed posts, including comment sections • Facebook Group (public) posts, including comment sections • Facebook Group (closed/private) posts, including comment sections • Facebook Channels	Yes	• Nilsimsa
	Messenger • Messenger (when E2EE not enabled) • Messenger Channels	Yes	• Nilsimsa
	Messenger • Messenger (when E2EE enabled)	No	
	Instagram • Instagram Feed • Instagram Direct (when E2EE not enabled) • Instagram Groups	Yes	• Nilsimsa
	Messenger • Instagram Direct (when E2EE enabled)	No	
	Threads • Threads	Yes	• Nilsimsa
Reddit	• Subreddits (public) • Subreddits (private)	Yes	• Snooron – Internal image hash-matching functionality • Rule-Executor-V2 (REV2) – automated enforcement system

	• Chat • Channels	No (but since implemented)	Implemented since reporting period • Snooron – Internal image hash-matching functionality • Rule-Executor-V2 (REV2) – automated enforcement system
WhatsApp	• Channels messages • Content in user reports	No	
Telegram	• Content in user reports	Yes	Internal Telegram Hash Matching System
	• Chats • Secret chats (user reports) • Group chats (public) • Group chats (private) • Channels (public) • Channels (private) • Stories	No	

Detecting new TVE

Providers were asked about the use of automated tools to proactively detect various forms of new or 'previously unknown' TVE. Hash-matching tools can only 'match' against previously identified and confirmed ('known') TVE and seek to prevent its ongoing dissemination. However, steps can also be taken to prevent the sharing of TVE when it is first created or shared, and before it has been identified and included in a database. There are technology options that enable service providers to proactively scan for this kind of 'first-generation' TVE.

For example, classifiers (**audio and/or visual classifiers**) are tools that use AI-powered pattern recognition to identify material or activity that is likely to depict or advocate TVE. These tools are trained on various datasets, including verified TVE, as well as material that does not contain TVE, in order to identify the markers of likely TVE. Depending on the datasets these classifiers have been trained on, they can be used to proactively scan and closely analyse images, videos, or written text to detect likely TVE. Providers can also scan text using natural language processing (NLP) tools which use machine learning to understand, analyse and moderate written language at scale.

Given these tools are identifying new, previously unknown material, which may rely on context or verification to confirm, this material is typically flagged for human review. When flagged for

human review, information about the final moderation decision can be fed back into the system to improve the accuracy of future automated detections.

eSafety asked about the use of technology to detect new TVE, in relation to sections 6(2) and 11 of the Determination.

Table 4: In response to the notices, the following information was given by service providers regarding the use of tools to identify new TVE images.

Service	Services/parts of services	Used tools to identify new TVE images	Names of tools used
Google	**YouTube** • YouTube profile pictures • YouTube video thumbnails	Yes	Proprietary Google image detection technology
	Drive • Drive (consumer version; stored content) • Drive (consumer version; content when it is shared)	No	
Meta	**Facebook** • Facebook Newsfeed posts, including comment sections • Facebook Group (public) posts, including comment sections • Facebook Group (closed/private) posts, including comment sections • Facebook Channels • Facebook Stories • Facebook profile pictures • Facebook Group profile pictures	Yes	Unified Content Model
	Messenger • Messenger Channels • Messenger Stories	Yes	Unified Content Model
	Messenger • Messenger (when E2EE not enabled) • Messenger (when E2EE enabled)	No	
	Instagram • Instagram Feed • Instagram Groups • Instagram Reels • Instagram profile picture • Instagram Groups profile picture	Yes	Unified Content Model

	Instagram • Instagram Direct (when E2EE not enabled) • Instagram Direct (when E2EE enabled)	No	
	Threads • Threads • Threads profile picture	Yes	Unified Content Model
Reddit	• Subreddits (public) • Subreddits (private) • Chat • Channels • Account profile pictures • Channel profile pictures	Yes	• Hive AI - AI image detection tooling; image optical character recognition (OCR) • Rule-Executor-V2 (REV2) – automated enforcement system
	• Subreddit profile pictures	Yes	• Hive AI - AI text detection tooling
WhatsApp	• Channels messages • Channels profile picture • Groups profile picture	Yes	• Whole Post Integrity Embeddings Service • CT Image Classifier
	• Content in user reports • Status • User profile picture • Communities profile picture	No	
Telegram	• Group chats (public) • Channels (public) • Stories • User profile picture • Group profile picture • Channel profile picture • Content in user reports	Yes	Internal Telegram AI and Machine Learning Models
	• Chats • Secret chats (user reports) • Group chats (private) • Channels (private)	No	

Table 5: In response to the notices, the following information was given by service providers regarding the use of tools to identify new TVE videos.

Service	Services/parts of services	Used tools to identify new TVE video	Names of tools used	Number of languages tools operated in
Google	**YouTube**	Yes	Proprietary Google classifier technology A	104
	Drive (consumer version; stored content)	No		
	Drive (consumer version; content when it is shared)	Yes	• Proprietary Google classifier technology A • Proprietary Google hashing technology	104
Meta	**Facebook** • Facebook Newsfeed posts, including comment sections • Facebook Group (public) posts, including comment sections • Facebook Group (closed/private) posts, including comment sections • Facebook Channels • Facebook Stories	Yes	Unified Content Model	101
	Messenger • Messenger Channels • Messenger Stories • Messenger Rooms			
	Instagram • Instagram Feed • Instagram Groups • Instagram Reels			
	Threads • Threads			
	Messenger • Messenger (when E2EE not enabled) • Messenger (when E2EE enabled)	No		
	Instagram • Instagram Direct (when E2EE not enabled) • Instagram Direct (when E2EE enabled)			

Reddit	• Subreddits (public) • Subreddits (private)	Yes	• Hive AI – video classification AI • Rule-Executor-V2 (REV2) – automated enforcement system • Google Vision OCR API – text detection	59
WhatsApp	• Channels messages	Yes	• Whole Post Integrity Embeddings Service	99
	• Content in user reports	Yes	• CT Text Classifier	99
	• Status	No		
Telegram	• Group chats (public) • Channels (public) • Stories • Content in user reports	Yes	Internal Telegram AI and Machine Learning Models	Telegram stated that it did not maintain a list of languages included in the training sets of its proactive detection tools and could not provide such a list in response to eSafety's questions in the Notice.
	• Chats • Secret chats (user reports) • Group chats (private) • Channels (private)	No		

Table 6: In response to the notices, the following information was given by providers regarding the use of tools to identify phrases, codes, hashtags indicating likely TVE in text.

Provider	Services/parts of services	Used tools to identify phrases, codes, hashtags indicating likely TVE in text	Names of tools used	Number of languages tools operated in
Google	**YouTube** • Username • Account description • Video titles • Video descriptions • Comments sections	Yes	BERT (Bidirectional Encoder Representations from Transformer)	104
	YouTube • Playlist titles	No		
	Drive • Drive (consumer version; stored content) • Drive (consumer version; content when it is shared)	No		
	Drive • Filename	Yes*	*Google clarified that there is no ongoing monitoring or scanning, but Google will scan for duplicates of known violative files on 'an ad-hoc or case by case basis'.	
Meta	**Facebook** • Facebook Newsfeed posts, including comment sections • Facebook Group (public) posts, including comment sections • Facebook Group (closed/private) posts, including comment sections • Facebook Channels • Facebook Stories • Facebook username • Facebook profile description	Yes	Unified content model	

	• Facebook group username (public and closed/private) • Facebook group profile description (public and closed/private)			101
	Messenger • Messenger Channels • Messenger Stories	Yes	Unified content model	
	Instagram • Instagram Feed • Instagram username • Instagram user bio • Instagram Groups • Instagram Groups username • Instagram Groups profile description • Instagram Reels	Yes	Unified content model	
	Threads • Threads • Threads Bio	Yes	Unified content model	
	Messenger • Messenger (when E2EE not enabled) • Messenger (when E2EE enabled)	No		
	Instagram • Instagram Direct (when E2EE not enabled) • Instagram Direct (when E2EE enabled)			
Reddit	• Subreddits (public) • Subreddits (private) • Chat • Channels	Yes	• Snooron – Keyword matching text classifier technology • Rule-Executor-V2 (REV2) – automated enforcement system • Hive AI - image optical character recognition (OCR)	27

	• Private messages • Account name • Account profile description • Subreddit name • Subreddit profile description	Yes	• Snooron – Keyword matching text classifier technology • Rule-Executor-V2 (REV2) – automated enforcement system	26
	• Channel name • Channel profile description • Subreddit wikis	No		
WhatsApp	• Channels messages • Channels profile description	Yes	• Whole Post Integrity Embeddings Service • CT Text Classifier	99
	• Content in user reports • Communities profile description • Groups profile description		• CT Text Classifier	99
	• Status • User profile description	No		
Telegram	• Group chats (public) • Channels (public) Stories • Profile username • Profile description • Group username • Group description • Channel username • Channel description • Content in user reports	Yes	Internal Telegram AI and Machine Learning Models	Telegram stated that it did not maintain a list of languages included in the training sets of its proactive detection tools and could not provide such a list in response to eSafety's questions in the Notice.
	• Chats • Secret chats (user reports) • Group chats (private) • Channels (private)	No		

Table 7: In response to the notices, the following information was given by providers regarding the percentage of reports sent for human review.

Provider	Services/parts of services	Percentage of user reports of TVE sent for human review	Percentage of TVE reports detected through automated tools (proactive detection) sent for human review
Google	YouTube	99%[31]	86.4%[32]
	Drive	100%	96%
Meta[33]	Facebook	83.4%	4.6%
	Messenger	39.7%	0.2%
	Instagram	87.8%	3.4%
	Threads	59.4%	3.2%
Reddit		100%[34]	66.5%[35]
WhatsApp		100%[36]	100%
Telegram		75%	65%

Blocking links to TVE

Service providers were asked about the use of proactive tools to detect and block URLs to TVE hosted on other platforms.

Experts in countering terrorism and violent extremism have warned that online extremists and pro-terror actors are increasingly attempting to avoid moderation on mainstream services by 'outlinking'[37] to TVE hosted on third-party platforms.

> 'Aware that their content can no longer achieve an enduring presence on the most well-known content sharing and social networking platforms, the strategy of [terrorist groups] for

[31] Google stated that this figure refers to 'videos uploaded from Australia'.
[32] Google stated that this figure refers to 'videos uploaded from Australia'.
[33] Meta noted that these figures represent Australian user data for the period 1 October 2023 to 29 February 2024.
[34] Reddit reported that the 100% refers to reports that users have made under its 'threatening violence' option and that Reddit has thereafter determined may be terrorist content.
[35] Reddit reported that the 66.5% refers to 'terrorist content' (as opposed to 'TVE') detected through automated tools that is sent for human review.
[36] WhatsApp provided the number of accounts that were banned or against which other enforcement actions were taken for TVE-related violations and which also had a user report over the last 30 days. WhatsApp stated that the data 'relates to user reports by Australian users' and is limited to the period 1 March to 30 April 2024 due to its data retention policies.
[37] An 'out-link' refers to a hyperlink that directs users from one website to an external website, serving as a digital pathway connecting one site to another.

> the past nine years has been to disseminate new propaganda via URLs that send the viewer to dozens of small platforms, on which the content is hosted'. TechUK[38]

Instead of directly posting text, images or videos of TVE (which may be more easily detected using proactive scanning tools), bad actors share links to platforms with weaker or non-existent moderation practices and policies[39] – including websites that are directly operated by designated TVE groups. Experts have also highlighted that TVE actors have exploited 'join-linking' – a feature on some messaging services that enables end-users to forward and share access to private groups – to promote and amplify groups devoted to TVE.[40] Such tactics are used by TVE actors and their sympathisers to disseminate pro-terror material and violent extremist propaganda, radicalise and recruit new adherents, and raise funds for terrorist activities.[41]

There are options available to service providers to detect and block URLs to TVE hosted on other platforms. For example, the non-profit organisation Tech Against Terrorism maintains a database of URLs known to be associated with TVE which it provides to participating industry members through an automated alert system.[42] The blocking of URLs is also a common practice across many online services for safety, security and legal reasons.

eSafety asked about measures to detect and block URLs to TVE, in relation to sections 6(2) and 11 of the Determination.

---

[38] techUK, How terrorists are capitalising on the cost of AI (Guest blog by Faculty), 16 Jan 2023, accessed 19 June 2024, URL: https://www.techuk.org/resource/natsec2023-faculty-16jan23.html

[39] Global Internet Forum to Counter Terrorism, 'Technical Approaches Output 1 – Gap Analysis and Recommendations', 2021, accessed 4 June 2024, URL: https://gifct.org/wp-content/uploads/2021/07/GIFCT-TAWG-2021.pdf; Organisation for Economic Cooperation and Development, 'Transparency reporting on terrorist and violent extremist content online 2022', 2022, accessed 4 June 2024, URL: https://gifct.org/wp-content/uploads/2021/07/GIFCT-TAWG-2021.pdf, https://www.oecd-ilibrary.org/science-and-technology/transparency-reporting-on-terrorist-and-violent-extremist-content-online-2022_a1621fc3-en

[40] Global Internet Forum to Counter Terrorism, 'Technical Approaches Output 1 – Gap Analysis and Recommendations', 2021, accessed 4 June 2024, URL: https://gifct.org/wp-content/uploads/2021/07/GIFCT-TAWG-2021.pdf; Middle East Media Research Institute, 'Pro-ISIS Telegram Channel Posts Links To WhatsApp Group Chat With Strict Religious Conditions For Joining', 2017, accessed 4 June 2024, URL: https://www.memri.org/cjlab/pro-isis-telegram-channel-posts-links-to-whatsapp-group-chat-with-strict-religious-conditions-for-joining

[41] Tech Against Terrorism, 'Report: The threat of terrorist and violent extremist operated websites', January 2022, accessed 4 June 2024, URL: https://techagainstterrorism.org/news/2022/01/28/report-the-threat-of-terrorist-and-violent-extremist-operated-websites

[42] Tech Against Terrorism, 'Terrorist Content Analytics Platform', 2024, accessed 4 June 2024, URL: https://techagainstterrorism.org/terrorist-content-analytics-platform

Table 8: In response to the notices, the following information was given by providers regarding blocking and source of URLs linking to known TVE hosted on other websites/services and 'join-links' to groups known to be associated with TVE.

Provider	Services/parts of services	Blocked URLs linking to known TVE hosted on other websites/services	Blocked 'join-links' to group chats associated with TVE	URL databases/sources used
Google	YouTube • Account description • Video descriptions • Comments sections	Yes	Yes	YouTube's own blocklist No external databases used
Meta	**Facebook** • Facebook Newsfeed posts, including comment sections • Facebook Group (public) posts, including comment sections • Facebook Group (closed/private) posts, including comment sections • Facebook Channels • Facebook profile description • Facebook Group profile description (public and closed/private)	Yes	Yes	Meta's 'own ongoing integrity work' and investigations by paid third party vendors
	Messenger • Messenger (when E2EE not enabled) • Messenger Channels	Yes	Yes	Meta's 'own ongoing integrity work' and investigations by paid third party vendors
	Messenger • Messenger (when E2EE enabled) • Messenger Rooms[43]	No	No	
	Instagram • Instagram Feed • Instagram Direct (when E2EE not enabled) • Instagram Bio	Yes	Yes	Meta's 'own ongoing integrity work' and investigations by

[43] Meta stated that it was not possible to share URLs in Messenger Rooms.

	• Instagram Groups • Instagram Groups profile description			paid third party vendors
	Instagram • Instagram Direct (when E2EE enabled)	No	No	
	Threads • Threads • Threads Bio	Yes	Yes	Meta's 'own ongoing integrity work' and investigations by paid third party vendors
Reddit	• Subreddits (public) • Subreddits (private) • Chat • Private messages • Channels • Account profile description • Subreddit profile description	Yes	Yes	• Reddit's own TVE hash list • Tech Against Terrorism (TAT) hash bank
	• Channel profile description • Subreddit wikis	No	No	
WhatsApp	• WhatsApp	No	No	
Telegram	• Chats • Secret chats (E2EE) • Group chats (public) • Group chats (private) • Channels (public) • Channels (private) • Profile description • Group description • Channel description	No	No	

Detecting TVE in livestreams and video calls

Terrorists can, and have, exploited live video to broadcast terror attacks on the internet. Terrorist attacks in Christchurch, Buffalo, and Halle demonstrate the way terrorists have weaponised livestreaming to amplify the effects of their violence. In the case of the 2019 Christchurch Mosque shootings, the perpetrator was able to broadcast his attack on Facebook

Live for 17 minutes before the livestream was discontinued.[44] In that time, approximately 200 people watched, from the terrorist's perspective, the murder of multiple people.[45] Five years on, recordings of this footage continue to be some of the most common TVE that Australians report to eSafety.

The immediate broadcast and subsequent circulation of this livestreamed content causes societal harms. It inflicts further pain and trauma on victims and their loved ones, helps bad actors glorify the actions of perpetrators, and advocates for or inspires copy-cat acts of violence against others. Detecting and rapidly removing livestreamed acts of terror is vital to ensure that bad actors cannot exploit online services to perpetrate these harms against society.

Detecting TVE in a live video is more technically challenging than detecting still images, given the volume of content transmitted. However, previous transparency reports published by eSafety have revealed that some companies have developed their own tools to detect livestreamed child sexual exploitation and abuse activity.

Other steps, such as prioritising human review of reports of livestreamed content, can also be taken by providers to reduce the likelihood of livestreamed TVE.

eSafety asked about the detection of livestreamed TVE, in relation to sections 6(2) and 11 of the Determination.

Table 9: In response to the notices, the following information was given by providers regarding the use of tools to detect TVE in livestreams or video calls.

Provider	Service	Measures in place to detect TVE in livestreams	Names of tools used	Interventions used (e.g., text classifiers, video classifiers, behavioural signals etc.)	Number of languages tools operated in
Google	**YouTube** • Livestream video • Live chat • Text associated with livestream (title and description)	Yes	Proprietary Google Classifier technology B	• Text classifiers • Video Classifiers • Audio Classifiers • Keyword detection	104

[44] Royal Commission of Inquiry into the Terrorist Attack on Christchurch Mosques on 15 March 2019', 'Report: Royal Commission of Inquiry into the terrorist attack on Christchurch masidjan on 15 March 2019', 2020, accessed 4 June 2024, URL: https://christchurchattack.royalcommission.nz/

[45] Royal Commission of Inquiry into the Terrorist Attack on Christchurch Mosques on 15 March 2019', 'Report: Royal Commission of Inquiry into the terrorist attack on Christchurch masidjan on 15 March 2019', 2020, accessed 4 June 2024, URL: https://christchurchattack.royalcommission.nz/

Meta	Facebook Live	Yes	• Proprietary Meta video hashing tool • Proprietary Meta Classifier 1 • Proprietary Meta Classifier 2	• Text classifiers • Video classifiers • Audio classifiers • Keywords • Behavioural signals	101
	Instagram Live				
	Messenger Rooms	No			
Reddit	n/a – Reddit does not have a livestream or video call function				
WhatsApp	Video calls	No			
Telegram	• Group video calls • Channel livestreams	No			

Percentage of proactive detection

The proportion of violative material that online services detect proactively is an indicator of the extent to which proactive detection tools are being deployed – or relied upon – by a service provider. A high percentage of proactive detections may indicate that services are highly reliant on automated tools to detect harmful or otherwise violative content. A low percentage of proactive detections may indicate that there has been limited deployment of automated tools and the service is instead more reliant on reports by users or trusted flaggers to identify and take action against rule violations.

eSafety asked about the percentage of TVE detected proactively, in relation to sections 6(2) and 11 of the Determination.

Table 10: In response to the notices, the following information was given by providers regarding the percentage of TVE detected proactively.

Provider	Services/parts of services	Percentage TVE proactively detected	Percentage TVE reported by users, trusted flaggers, other
Google	YouTube	95.3%[46]	0.8% Priority Flaggers[47] 3.9% users[48]
	Drive (consumer version)[49]	~66%	34%
Meta[50]	Facebook Newsfeed	96.2%	3.8%
	Facebook Groups (Public)	89.9%	10.1%
	Facebook Groups (Closed/Private)	93.3%	6.7%
	Messenger (E2EE and when E2EE not enabled)	100%	0%
	Instagram Feed	99.4%	0.6%
	Instagram Direct (E2EE and when E2EE not enabled)	100%	0%
	Threads	93.2%	6.8%
Reddit[51]	• Subreddits (public)	79.4%	20.6%
	• Subreddits (private)	100%	0%
	• Chat • Private messages • Channels • Subreddit wikis	Reddit reported that during the report period it did not have any terrorism-related removals in these parts of the service	
WhatsApp[52]		91%[53]	9%[54]
Telegram	Chats	N/A	100%
	Secret Chats (E2EE)	N/A	100%
	Group chats (public)	67%	33%
	Group chats (private)	82%	18%

[46] Google stated that this figure 'represents the percentage of videos that were uploaded from Australia'
[47] Google stated that this figure 'represents the percentage of videos that were uploaded from Australia'
[48] Google stated that this figure 'represents the percentage of videos that were uploaded from Australia'
[49] Google stated that due to its data retention policies, some of the data requested by eSafety was no longer available and that these figures were calculations based on 'good-faith efforts and the 'best data that is currently available for the Reporting Period'.
[50] Meta noted that these figures represent content created by Australian users that was removed due to TVE policy violations during the period 1 October 2023 to 29 February 2024.
[51] Reddit stated that when it actions content under its 'violence policy' it categorises those removals either under the 'broader violence category' or the 'narrower terrorism sub-subcategory' not as 'TVE'.
[52] WhatsApp stated that these figures represent TVE created by Australian users during the report period.
[53] For percentage of TVE 'proactively detected' WhatsApp reported on instances where it did not receive a report against the relevant account in the 30 days prior to enforcement.
[54] For percentage of TVE 'reported by users, trusted flaggers or other' WhatsApp reported on instances where it did receive a report against the relevant account in the 30 days prior to enforcement.

	Channels (public)	69%	31%
	Channels (private)	79%	21%
	Voice and video calls (public and private)[55]	N/A	N/A
	Group video calls (public and private)[56]	'Included in group chats'	
	Stories	60%	40%

User reporting

Reporting and complaints mechanisms enable users to flag and alert online service providers to specific material and activity that is illegal, harmful or otherwise in breach of a service's terms of service.

The importance of user reporting as a safety measure is reflected in the following Expectations:

- Section 13: that providers have clear and readily identifiable mechanisms that enable end-users to report, and make complaints about, certain material (including forms of TVE material)
- Section 14(1)(c): that providers have policies and procedures for dealing with reports and complaints mentioned in section 13 or 15
- Section 15(1) and (2): that providers have clear and readily identifiable mechanisms that enable end-users, and those ordinarily resident in Australia, to report, and make complaints about, breaches of the service's terms of use.

eSafety has published regulatory guidance for the Expectations,[57] setting out the expectations of the online industry regarding the provision of mechanisms for users to report and make complaints.

The regulatory guidance sets out that a reporting or complaint mechanism is likely to be 'clear' if users are presented with categories that describe the issue they wish to report. Issue-specific reporting options allow services to prioritise user reports for rapid response and escalation depending on their severity. In circumstances where online content or conduct represents a serious threat to life, health or safety – such as an unfolding terrorist attack being broadcast over a livestream – issue-specific user reporting is imperative. This is because it helps ensure

[55] In answer to a follow-up question from eSafety to clarify why its answer was 'N/A' for voice and video calls Telegram stated that voice and video calls could not be directly reported by end-users using in-service reporting tools. Instead, 'calls are reported together with their respective community (via the community info section and by additionally including a subset of objectionable sample messages)'.
[56] Telegram stated that its video group call data was included in the relevant group chat statistics because 'information on resulting bans is not stored separately'.
[57] eSafety.gov.au, Basic Online Safety Expectations Regulatory Guidance, accessed 12 February 2025, URL: https://www.esafety.gov.au/industry/regulatory-guidance

that the report is given the necessary prioritisation to enable moderators to disrupt the harm, remove the offending material and report the incident to law enforcement agencies as quickly as possible, where appropriate.

The regulatory guidance also sets out that a reporting or complaint mechanism is 'readily identifiable' if it can be quickly and easily accessed by an individual without barriers, at every part of the user experience. For example, reporting and complaints mechanisms should be provided on all aspects of a service so that an individual can report all relevant material and activity – including material they have seen in a post, a livestream, a video chat or direct communication, or activity by another end-user or by a group or forum. These mechanisms should be consistently accessible for individuals whether the service has been accessed via an app or browser, and they should be available to all users, regardless of whether they are logged into an account or not.

In addition to making reporting mechanisms available to ordinary users, many online services also use 'trusted flagger' programs or other specialised reporting avenues that enable qualified subject matter experts, government agencies and law enforcement agencies to refer certain material or activity for review through expedited escalation pathways. These 'Trusted Flagger' pathways enable content moderators to prioritise reports that carry a higher expectation of legitimacy and, in some cases, that may relate to an imminent threat to life or health.

eSafety asked providers about the steps taken to implement user reporting and complaints mechanisms on their services, in relation to sections 13, 14 and 15 of the Determination.

Table 11: In response to the notices, the following information was given by providers regarding in-service and expert/authority reports of TVE.

Service	Services/parts of services	Ability for end-users to report instances of TVE in-service	Category used to report TVE in-service	Separate reporting mechanism for experts and authorities to report to provider	Separate reporting mechanism available for following entities
Google	**YouTube**	Yes	• Promotes terrorism; or • Hateful or abusive content; or • Violent or repulsive content	Yes	• Law enforcement • Trusted flaggers • Regulatory or other public authorities
	Drive (consumer version; content when it is shared)	Yes	• Violent organisations and movements content; or		

			• Violence; or • Hate Speech		
Meta	**Facebook** • Facebook Newsfeed • Facebook Groups • Facebook Stories	Yes	• Terrorism	Yes	• Law enforcement • Trusted flaggers • Regulatory or other public authorities • Civil society groups
	Facebook • Facebook Channel		• Sharing Inappropriate Things -> Violent or Graphic content		
	Messenger • Messenger (when E2EE enabled) • Messenger (when E2EE not enabled) • Messenger Channels	Yes	• Sharing Inappropriate Things -> Violent or Graphic content	N/A[58]	
	Messenger • Messenger Stories	Yes	• Violence	N/A	
	Messenger • Messenger Rooms	No			
	Instagram • Instagram Feed • Instagram Direct (when E2EE enabled) • Instagram Direct (when E2EE not enabled) • Instagram Groups • Instagram Reels	Yes	• Violence or dangerous organisations	Yes	• Law enforcement • Trusted flaggers • Regulatory or other public authorities • Civil society groups
	Threads • Threads	Yes	• Violence or dangerous organisations	N/A	
Reddit[59]	• Subreddits (public and private) • Chat • Private Messages • Channels	Yes	• Threatening violence	Yes	• Law enforcement • Trusted flaggers

[58] Meta was not asked about separate reporting mechanisms for experts and authorities on Messenger and Threads.
[59] Reddit stated that when it actions content under its 'violence policy' it categorises those removals either under the 'broader violence category' or the 'narrower terrorism sub-subcategory' not as 'TVE'.

	• Subreddit Wikis	No	N/A		• Regulatory or other public authorities
WhatsApp	• Direct Messages (including Groups) • Communities • Channels • Status	Yes	• Report	Yes	• Law enforcement • Trusted flaggers • Regulatory or other public authorities • Civil society groups
Telegram	• Chats • Secret Chats	Yes	• Block user > Report Spam[60]	Yes	• Law enforcement • Trusted Flaggers • Regulatory or other public authorities • 'International organizations'
	• Group chats (public) • Group chats (private) • Channels (public) • Channels (private) • Stories	Yes	• Violence		
	• Voice calls • Video calls	No[61]	N/A		

Human moderation, expertise and resources

Human moderation refers to the practice of employing human beings to assess whether users of an online service are abiding by its terms of service. This may involve human moderators actively monitoring a service and taking proactive action when they identify content or activity that breaches a service's terms of service. It may also involve human moderators responding to reports submitted by users and trusted flaggers or when material or activity is flagged for human review by automated tools. Human moderators are typically employees or contractors employed by a service provider. Such moderators should be trained in how to assess and minimise suspected violations, interpret relevant rules and policies (taking into account any relevant context) and take action consistent with the service's policies.

[60] Telegram subsequently clarified that the 'Bock + Report Spam' reporting flow is only available when the Chat or Secret Chat is 'initiated by non-contacts and strangers'. eSafety understands that when an end-user wishes to report a message from an account they have already added as a contact, the only option in-service is to 'Block user'. See section 4A of Telegram's summary for further details.

[61] Telegram's original response to the Notice stated that end-users could make in-service reports about voice calls and video calls using a 'Violence (via the community info section)' reporting category. In response to a follow-up question from eSafety, Telegram subsequently stated that in-service reporting for voice and video calls was not available during the report period. Instead, Telegram stated that 'calls are reported together with their respective community (via the community info section and by additionally including a subset of objectionable sample messages)'.

Human moderation is particularly useful in circumstances where the facts of a suspected harm or violation are unclear or context-dependent and require review by staff capable of understanding context and the nuances of speech, behaviour and culture. For example, some TVE, such as recordings of terrorist attacks or excerpts from terrorist manifestos, may be shared online for legitimate journalistic or academic reasons – or they may be shared by sympathetic violent extremists as a form of hateful pro-terror propaganda. In such cases, automated moderation tools may be insufficient for determining the intent of the material and the appropriate moderation response. For this reason, service providers may elect to use automated tools in conjunction with human moderators. Automated tools may proactively detect a suspected violation and flag it for review. A human moderator would then assess and make the ultimate moderation decision.

eSafety asked providers to report on the use of human moderation to detect and address TVE on their services.

Languages moderators operated across

Assessing complex, context-dependent harms requires linguistic, regional and cultural understanding. There is a risk of losing important nuance when proactive detection measures operate in a small number of languages and there is reliance on language translation tools.[62] For this reason, it is particularly important that services have human moderators operating in the languages of the communities they offer services to.

The top five languages other than English spoken in Australian homes are Arabic, Cantonese, Mandarin, Vietnamese and Punjabi.[63]

eSafety asked about the languages human moderators operated across, in relation to sections 6, 11, 13, 14, and 15 of the Determination.

[62] Details covering the languages supported by proactive detection tools used to detect suspected TVE can be found in the section 'Proactive detection'. A full list of the languages human moderators operated across as well as languages supported by proactive detection tools can be found in each provider summary at the end of this report.
[63] Australian Bureau of Statistics, 'Cultural diversity: Census', 28 June 2021, URL: https://www.abs.gov.au/statistics/people/people-and-communities/cultural-diversity-census/latest-release#:~:text=Top%205%20languages%20used%20at,Punjabi%20(0.9%20per%20cent).

Table 12: In response to the notices, the following information was given by providers regarding the languages human moderators operated across.

Service	Number of languages		Languages	
	Employees	Contractors	Employees	Contractors
Google YouTube Drive Gemini	1	80	English	Afrikaans; Amharic; Arabic; Azerbaijani; Belarusian; Bengali; Bosnian; Bulgarian; Burmese; Cantonese; Croatian; Czech; Danish; Dutch; English; Estonian; Ethiopian-Amharic; Ethiopian-Oromo; Ethiopian-Tigriniya; Filipino; Finnish; French; German; Greek; Gujarati; Hausa; Hebrew; Hindi; Hungarian; Igbo; Indian Languages; Indonesian; Irish; Italian; Japanese; Kazakh; Khmer; Korean; Kurdish; Laos; Latvian; Lithuanian; Macedonian; Malay; Malayalam; Mandarin; Mandarin/Cantonese; Marathi; Norwegian; Oriya; Oromo; Pashto; Persian; Polish; Portuguese; Portuguese-BR; Punjabi; Romanian; Russian; Serbian; Sinhalese; Slovenian; Somali; Spanish; Swahili; Swedish; Tagalog; Tajik; Tamil; Telugu; Thai; Tigrinya; Turkish; Ukrainian; Urdu; Uyghur; Uzbek; Vietnamese; Yoruba; Zulu
Meta Facebook Messenger Instagram Threads	89	84	Albanian; Amharic; Arabic; Arabic (Gulf); Arabic (Levant, Egypt, Iraq); Arabic (Sudan); Armenian; Assamese; Azerbaijani; Bambara; Belarusian; Bemba; Bengali; Bengali (India); Bosnian; Bulgarian; Burmese; Cantonese; Croatian; Czech and Slovak; Danish; Dari; Dutch; English; Estonian; Filipino; French; French (Sub-Saharan Africa); Fula; Georgian; German; Greek; Gujarati; Hausa; Hebrew; Hindi; Hungarian; Igbo;	Afrikaans; Albanian; Amharic; Arabic; Armenian; Assamese; Azerbaijani; Bengali; Bhojpuri; Bosnian; Bulgarian; Burmese; Cantonese; Chhattisgarhi; Czech; Croatian; Danish; Dari; Dhivehi; Dutch; English; Estonian; Finnish; French; Ganda; Georgian; German; Greek; Gujarati; Hausa; Hebrew; Hindi; Hungarian; Indonesian; Italian; Japanese; Kannada; Kazakh; Khmer; Konkani; Korean; Kurdish; Lao; Latvian; Lithuanian; Luganda; Malay; Malayalam; Maltese; Mandarin; Marathi; Marwari; Meitei; Mizo; Mongolian; Nepali; Oriya; Oromo; Pashto/Pushto; Persian; Polish; Portuguese; Punjabi;

			Indonesian; Italian; Japanese; Kannada; Kazakh; Khmer; Kirundi; Kituba; Korean; Kurdish; Latvian; Lingala; Lithuanian; Maghreb Arabic; Malay; Malayalam; Mandarin; Marathi; Mauritian Creole; Mongolian; Nepali; Norwegian; Oriya; Oromo; Pashto/Pushto; Persian; Polish; Portuguese; Punjabi; Romanian; Russian; Serbian; Sindhi (India); Sindhi (Pakistan); Sinhala; Somali; Spanish (Latin America); Spanish (Spain); Swahili; Swedish; Tamil; Telugu; Thai; Tigrinya; Turkish; Ukrainian; Urdu (India); Urdu (Pakistan); Vietnamese; Yoruba; Zulu	Romanian; Russian; Serbian; Sindhi; Sinhala; Somali; Spanish (Castilian); Swahili; Swedish; Tagalog; Tamil; Telugu; Thai; Tigrinya; Tulu; Turkish; Ukrainian; Urdu; Uzbek Vietnamese; Zulu
Reddit	13	8	English; French; Spanish; Portuguese; Arabic; Russian; German; Turkish; Urdu; Hindi; Telugu; Shona; Zulu	English; French; Spanish; Portuguese; Russian; Turkish; Hindi; German
WhatsApp	N/A	6	WhatsApp stated that it 'relies on the language capabilities of its human review teams, who are contractors' WhatsApp subsequently stated that it 'provides its reviewers with translation tools to enable them to review material in languages other than their native languages.'	English; Spanish; Arabic; Urdu; Pashto; Farsi

Telegram[64]	N/A[65]	47	N/A	Amharic; Arabic; Azerbaijani; Bulgarian; Chinese (traditional and simplified); Croatian; Czech; Danish; Estonian; Farsi; Filipino; Finnish; French; Georgian; German; Greek; Hindi; Icelandic; Indonesian; Italian; Japanese; Kazakh; Korean; Kyrgyz; Luganda; Lunyakore; Lusoga; Malay; Moldavian; Norwegian; Polish; Portuguese (Brazil); Portuguese (Europe); Romanian; Russian; Serbian; Shona; Spanish; Swahili; Swedish; Tajik; Turkish; Ukrainian; Urdu; Uzbek; Yoruba

Dedicated teams to minimise TVE

Specialist teams with the relevant training in a particular form of online harm are well placed to more effectively and efficiently reach an informed, appropriate and timely moderation decision when triaging complex or high-risk cases, thus enhancing the overall safety on a service.

In the case of a livestreamed terrorist attack, the content may present such a clear and widespread threat of online harm that it requires an accelerated escalation pathway to dedicated crisis-response staff who have the skills and authority to take immediate action.

Dedicated TVE teams are also better positioned to anticipate and recognise trends and changes in an online landscape and are able to give informed, iterative feedback that strengthens the policies and processes used by providers to safeguard their services.

eSafety asked service providers about dedicated trust and safety teams responsible for minimising TVE on their services, in relation to sections 6 and 11 of the Determination.

[64] Telegram also advised that since the report period, it had expanded the languages covered by its contracted content moderators by adding Afrikaans, Bengali (Bangladesh), Chichewa (Zambia), Dhivehi (Maldives), Dutch, Gujarati, Kabyle (Algeria), Kinyarwanda, Lithuanian, Macedonian, Punjabi, Sinhalese (Sri Lanka), and Thai.
[65] Telegram stated that 'all ordinary moderators' on Telegram are contractors.

Table 13: In response to the notices, the following information was given by providers regarding dedicated trust and safety team(s) responsible for minimising TVE.

Provider	Dedicated trust and safety team responsible for minimising TVE	Number of employees	Number of contractors	Surge team(s) to respond to TVE crisis	Number of employees	Number of contractors
Google YouTube	No[66]			No[67]		
Meta Facebook Instagram	Yes	10	1	Yes	Cross-team rapid response protocol with 24/7 coverage by on-call employees	
Reddit	Yes	26	120	Yes	35	1
WhatsApp	Yes	6	0	Yes	24/7 escalation coverage by on-call employees Size of surge team depends on nature of event.	
Telegram[68]	Yes	4	0	Yes	3	13

Median time to respond to user reports

Measuring the median time taken to reach a content moderation outcome in response to a user report about TVE gives service providers insight into the efficacy of their trust and safety systems and resources and helps track improvements over time. A lengthy median response time may indicate that a service provider's trust and safety systems and processes are under-resourced or not optimally calibrated to respond to a particular type of harm in the most efficient way. For material and activity like TVE, which has the potential to cause significant harm, it is particularly important that service providers have systems and processes in place that enable them to review user reports and take relevant action as soon as possible to minimise harm to users on their services, and the wider public.

Service providers were instructed to calculate this metric from the time a user report was made to the time of a content moderation outcome or decision (such as removing the content,

[66] Google stated that 'As of 31 December, 2023, YouTube had 3,455 humans evaluating content in English, and 9,813 humans conducting language agnostic reviews.

[67] Google stated that YouTube has 'rapid response capabilities' to ensure that it responds to major incidents, including livestreamed terrorist attacks.

[68] Telegram stated that these figures were specific to 'staff that may from time to time be involved in decisions regarding content or reports from Australia and do not reflect or approximate the total number of global trust and safety personnel contracted by Telegram'.

banning the account, or deciding that no action should be taken). The methods used to calculate these figures differed, and each service provider's methodology is outlined in their provider-specific summary.

eSafety asked providers about the median time taken to respond to user reports about TVE, in relation to sections 6, 8, 11, and 14 of the Determination.

Table 14: In response to the notices, the following information was given by providers regarding the median time taken to reach an outcome after receiving a user report about TVE.

Provider	Services/parts of services	Median time taken to reach an outcome after receiving a user report about TVE	
		Reports from users globally	**Reports from users in Australia**
Google	**YouTube**[69]	4.4hrs[70]	Google reported that this information was not available
	Drive (consumer version; content when it is shared)	10.2hrs	2.9hrs
Meta[71]	Facebook Newsfeed	6.5 hours	4.2 hours
	Facebook Group (public)	6.7 hours	2.5 hours
	Facebook Groups (closed/private)	0.8 hours	2 hours
	Messenger (when E2EE enabled)	0.1 hours	0.1 hours
	Messenger (when E2EE not enabled)	0.1 hours	0.1 hours
	Instagram Feed	24.4 hours	15.5 hours
	Instagram Direct (when E2EE enabled)	4.3 hours	Meta reported that it did not have any reports from Australian users where content was determined to violate TVE policies

[69] Google reported that YouTube's figures were based on data that is not TVE-specific and were from outside the report period. Google stated that YouTube did not have data to distinguish the median time to enforce user flags based on country of origin or specific to its TVE policies. Following a request for clarification by eSafety, Google stated that the data is based on a study completed in July 2022 and that it relates to user flags on videos that are potentially violative of community guidelines, including guidelines related to TVE.

[70] Google reported this figure as '15 min for automated review of the flag' and 'Approx 4.4 hours for flags referred for human review'.

[71] Meta noted that these figures represent data from 1 October 2023 to 29 February 2024. Meta also reported that it does not ordinarily track or report data regarding response times to user reports that differentiates when E2EE is and is not enabled on Messenger and Instagram Direct. Meta stated the data provided for these surfaces was 'sourced from non-core datasets and cannot be verified or validated'. It added that 'while Meta has sought to provide accurate data to the best of its ability, Meta has material concerns about the reliability of this data and considers that this data is not sufficiently robust to be used for further analysis.'

	Instagram Direct (when E2EE not enabled)	5.8 hours	3 hours
	Threads	56.3 hours	59.5 hours
Reddit	Subreddits (public)[72]	62.2 hours[73]	31.3 hours[74]
WhatsApp[75]	Direct messages (including Groups)	25.3 hours	24.13 hours[76]
	Communities	24.8 hours	WhatsApp reported 'no Reported TVE Accounts for Communities' in report period
	Channels	24.5 hours	25.3 hours[77]
Telegram	• Chats • Secret Chats	18 hours[78]	18 hours[79]
	• Group chats (public) • Group chats (private) • Channels (public) • Channels (private)	15 hours[80]	15 hours[81]

[72] Reddit reported that there were no user reports confirmed to be terrorist content on the other parts of its service queried by eSafety during the report period.
[73] Reddit noted that users may report material that may be terrorist and/or violent extremist material under the violence reporting option, or potentially under the hate reporting option. Reddit further noted that it has no way to distinguish a user report of TVE from non-TVE violations of these rules, and that it therefore does not have data on the median time taken to reach an outcome after receiving "user reports of TVE" on the service. Reddit also noted that reports that its human safety team determines may relate to terrorist content are sent to a specialised terrorism queue for further human review. The data presented in this table is the median time between a user report and ticket closure for reports escalated to Reddit's specialized terrorism queue.
[74] See footnote above.
[75] WhatsApp reported that these figures reflect enforcement action taken against accounts that were banned for TVE-related violations and had also received a user report over the past 30 days. WhatsApp stated that due to the absence of issue-specific reporting options, WhatsApp cannot identify user reports where the user intended to report TVE specifically. WhatsApp also stated that because it does not log enforcement actions against specific user reports, it was 'not possible … to calculate the median time taken to reach an outcome after receiving a user report of TVE with precision.' WhatsApp reported that these figures are based on the assumption that the 'maximum amount of time' between the user report being made and it being 'enqueued for human review is 24 hours' plus the addition of the time then taken for enforcement action for each service.
[76] WhatsApp reported that it stores data related to Australian users for rolling 90-day periods. The information relating to reports from Australian users is limited to the period 9 February 2024 – 8 May 2024 and relates to a total of 4 user reports.
[77] WhatsApp reported that it stores data related to Australian users for rolling 90-day periods. The information relating to reports from Australian users is limited to the period 9 February 2024 – 8 May 2024 and relates to a total of 4 users.
[78] Telegram stated that to calculate these figures it registered the net time frames between the submission of each report and the moderator's decision with respect to that report.
[79] Telegram stated that it 'currently doesn't have the technical means to provide separate statistics by country'.
[80] Telegram stated that to calculate these figures it registered the net time frames between the submission of each report and the moderator's decision with respect to that report.
[81] Telegram stated that it 'currently doesn't have the technical means to provide separate statistics by country'.

Staffing levels

In 2023, both Google and Meta announced reductions to their staffing numbers.[82] The respective announcements did not disclose how the reductions would impact the resourcing of trust and safety functions on their services. Resourcing of trust and safety teams is important for ensuring online safety. Based on eSafety's observations over the past nine years of online safety regulation, companies with low numbers of trust and safety personnel may have reduced capacity to respond to TVE and other online harms.

eSafety asked Meta and Google to report on how their respective staffing levels for content moderators and other trust and safety personnel changed, in relation to sections 6 and 11 of the Determination.

Table 15: In response to the notices, the following information was given by Meta and Google regarding changes in trust and safety staffing levels.

Category of staff	Google YouTube, Drive, Gemini		Staff change by percentage	Meta[83]		Staff change by percentage
	1 April 2023	29 February 2024		31 March 2023	31 December 2023	
Engineers employed by service focussed on trust and safety	1305	1294	-0.8%	1,862	1,814	-2.6%
Content moderators employed by service	316	341	+7.9%	0[84]	0	N/A
Content moderators contracted by service	39,606	39,552	-0.1%	28,965	25,905	-10.6%

[82] Google, 'A difficult decision to set us up for the future', 20 Jan 2023, accessed 4 June 2024, URL: https://blog.google/inside-google/message-ceo/january-update/ ; Facebook, 'Update on Meta's year of efficiency', 14 March 2023, accessed 4 June 2024, URL: https://about.fb.com/news/2023/03/mark-zuckerberg-meta-year-of-efficiency/

[83] Meta reported that it could not provide staff data specific to the dates specified in the notice because it runs reports on its organisational numbers on a quarterly basis. Meta provided data as at 31 March 2023 and 31 December 2023 as an alternative.

[84] Meta reported that 'content moderators are generally employed by Meta's vendors'. Meta further reported that at 31 March 2023 there were 3,159 employees in its 'global operations team' and as at 31 December 2023 the figure was 1,967. Meta stated that its 'global operations team' focuses on 'work related to content moderation work (e.g., quality reviews, building protocols, managing contractors etc)'

Trust and safety staff employed (other than engineers and content moderators)	1,416	1,265	-10.7%	5,265[85]	3,803	-27.8%

Table 16: In response to the notices, the following information was given by Reddit, WhatsApp and Telegram regarding trust and safety staffing levels (both employed and contracted).

Service	Engineers employed by provider focussed on trust and safety		Content moderators employed by provider	Content moderators contracted by provider	Trust and safety staff employed by providers (other than engineers and content moderators)	
	#employees	#contractors	#employees	#contractors	#employees	#contractors
Reddit	82	7	15	107	71	23
Total[86]	89		122		94	
WhatsApp[87]	117		0[88]	1,365	266[89]	
Telegram[90]	5		0	150	4	

Volunteer or 'community' moderation

Volunteer or 'community' moderation is a model of content moderation where responsibility for enforcing community rules and regulating content is, at least partially, given to users who volunteer for the role. Depending on the particular service, volunteer moderators may have the ability to create new groups or forums, accept and remove members and establish additional rules and norms that apply in those communities. They may also have the ability to enforce service-wide policies and community-specific rules with a range of moderation tools. Volunteer moderators can be appointed by the service, self-appointed, or appointed by the specific groups that they operate in.

---

[85] Meta reported that this cohort included employees 'working in global operations and other non-engineering tech functions (i.e., product managers, researchers, designers, etc), legal, and policy'.
[86] Reddit noted that as of 29 February 2024, the total number of Reddit employees was 2030 and the total number of Reddit contractors was 989.
[87] WhatsApp reported that these figures were WhatsApp/Meta numbers focussed on WhatsApp as at 31 December 2023
[88] WhatsApp stated there are 'Nil' content moderators employed by WhatsApp, and that 'content reviewers are generally employed by Meta's vendors'. WhatsApp further stated that there were 'around 208 employees' focused on WhatsApp in WhatsApp/Meta's global operations team, which focuses on 'work related to review of content (e.g., quality reviews, building protocols, managing contractors etc)'.
[89] WhatsApp reported that this cohort included employees 'working in global operations and other non-engineering tech functions (i.e., product managers, researchers, designers, etc').
[90] Telegram stated that these figures represented the number of staff who 'may from time to time be involved in decisions regarding content or reports from Australia and do not reflect or approximate the total number of global content moderation and trust and safety personnel contracted by Telegram.' Telegram also stated that Australian end-users make up less than 0.2% of its monthly active users.

Community moderation has many legitimate uses and benefits. However, the model also carries the inherent risk of enabling bad actors to create communities that promote and legitimise terrorist and other extremist ideologies. The risk that community moderation may facilitate illegal, harmful or otherwise violative material and activity is elevated when a service enables volunteer moderators to create and manage 'closed groups' where activity inside the community is shielded from public view.

Closed groups are user-created forums or groups on a service that are only visible and accessible to their approved members. These closed groups, by design, inhibit public insight into the kinds of content being shared and the activity occurring inside them. The contents of closed groups are invisible to other end-users of the service who have not been accepted as members. Groups protected by end-to-end encryption are also inaccessible to trust and safety staff and automated detection tools. As a result of this reduced public insight, volunteer moderators have a heightened level of autonomy in setting standards and enforcing community rules. This creates risks that closed environments may be exploited as spaces where illegal, harmful and otherwise violative material and activity is understood to be permitted, or even actively encouraged.

Also, where there is a lack of engagement between volunteer moderators and the Trust and Safety staff of a service there is an increased risk of bad actors continuing to offend, because a volunteer moderator may only ban an offender from a specific channel or group, rather than the whole service.

eSafety asked service providers to report on volunteer moderation on their services and the tools, policies and processes they have in place to ensure that volunteer moderators are setting and enforcing appropriate safety standards.

eSafety asked providers about the systems and processes they have in place for volunteer moderators, in relation to sections 6, 11, and 14 of the Determination.

Table 17: In response to the notices, the following information was given by providers about systems and processes in place to ensure the setting and enforcing of appropriate safety standards by volunteer moderators.

Service	Standards policy, or similar, outlining volunteer moderator responsibilities and expectations	Ability for end-users to report in-service when volunteer moderators not meeting required responsibilities	Professional trust and safety staff automatically notified of an account removal by a volunteer moderator for TVE violation
Meta (Facebook)	Yes	Yes[91]	No
Reddit	Yes	No[92]	No[93]
WhatsApp	No[94]	Yes[95]	No
Telegram	Yes[96]	Yes[97]	No[98]

Preventing recidivism

In an online safety context, recidivism refers to banned or suspended users re-registering to an online service with new details to continue perpetrating harm. This can take the form of multiple fake accounts, including automated accounts or bots.

[91] Meta responded 'yes'. Meta's response indicated that a user can report the group in-service, it did not indicate that a specific report about a volunteer moderator can be made in service.
[92] Reddit reported that users may report violations of the Moderator Code of Conduct using a form on the Help Centre.
[93] Reddit responded 'Yes' that trust and safety staff are informed when a volunteer moderator removes an account from subreddits and/or channels (both public and private) for TVE breaches. Reddit reported that user reports of policy breaches go to both the moderation teams of the subreddit where the content was posted and to Reddit and therefore that Reddit will already be aware of any content removed by a volunteer moderator as a result of a user report. Following a subsequent question from eSafety, Reddit reported that it's trust and safety staff are not automatically informed when a volunteer moderator removes an account from a subreddit or chat channel.
[94] WhatsApp reported that 'responsibility for enforcing WhatsApp's policies remains with WhatsApp. Community admins are, like all WhatsApp users, encouraged to report behavior or content that may violate WhatsApp's Terms of Service to WhatsApp.'
[95] WhatsApp stated that end-users are able to report a Community via in-service reporting tools. WhatsApp qualified that this does not necessarily allow reporting of the Community admin personally
[96] Telegram initially reported 'no' and stated that it 'relies on contracted professional moderators. It did not have volunteer moderators as at 29 February 2024 and does not to date'. Following consultation with Telegram on the proposed report for publication, Telegram noted that it had interpreted eSafety's definition of 'volunteer moderator' differently and updated its response.
[97] Telegram responded 'Yes'. Telegram's response indicated that a user can report the Community in-service. It did not indicate that a specific report about a volunteer moderator can be made in-service.
[98] Telegram responded 'Yes' that trust and safety staff are informed when a volunteer moderator removes an account from a public channel, private channel or group for TVE breaches. Telegram's response stated that its administrators 'may' opt to report the removal of 'a user or their messages (in whole or in part) from a group' to Telegram with a detailed description of the infringement. eSafety understands that Telegram trust and safety are therefore not automatically informed when a volunteer moderator removes an account.

Where a service provider operates more than one service, preventing recidivism can involve ensuring that bad actors banned on one service are also banned on its other services.

Detecting recidivism

eSafety asked providers to report on the measures used to detect and prevent recidivism for TVE-related breaches on their services.

As with previous transparency reports, eSafety has chosen not to publish the specific indicators reported by service providers to prevent recidivism, to avoid this information being misused by bad actors. Instead eSafety has sought to demonstrate the range of indicators used. This is an imprecise metric, as some indicators were more important than others and some service providers used certain indicators more proactively and rigorously than others. However, eSafety's view is that, in general, service providers that are looking for a wider range of indicators to detect recidivism will have a better chance of preventing the re-registration of banned users.

eSafety uses the following terms to give an impression of the extent of the indicators used by services in the following table:

- Minimal: a small number
- Several: a moderate number
- Multiple: a significant number

eSafety asked providers about the signals and indicators used to prevent recidivism on their services, in relation to sections 6(2), 9, 11 and 14(2) of the Determination.

Table 18: In response to the notices, the following information was given by providers about the steps and indicators taken on their services to prevent recidivism.

Provider	Parts of service	Used steps to prevent recidivism	Number of indicators
Google	YouTube	Yes	Multiple
	Drive	Yes	Minimal
Meta	Facebook	Yes	Multiple
	Messenger	Yes	Multiple
	Instagram	Yes	Multiple
	Threads	Yes	Multiple
Reddit		Yes	Multiple
WhatsApp		Yes	Minimal
Telegram		Yes	Minimal

Preventing banned groups from being recreated

The ability of users to create groups and channels dedicated to the promotion and legitimisation of terrorist and extremist ideologies poses particular risks for the spread of TVE on a service. Bad actors can create online spaces where terrorist and violent extremist rhetoric is allowed or actively encouraged. This can lead to 'echo chambers' where harmful ideologies are unchecked by dissenting views and therefore have a radicalising effect on users.

Detecting and deactivating groups and channels devoted to terrorism and violent extremism is an important intervention that service providers can take to prevent networks of terrorists and violent extremists from becoming embedded on their services. Preventing such groups from being recreated is equally important for enforcing service bans and maintaining resistance against any bad actors attempting to return to the platform.

eSafety asked service providers about steps taken to prevent banned groups from being recreated after they have been banned for TVE-related violations, in relation to sections 6, 11, and 14 of the Determination.

Table 19: In response to the notices, the following information was given by providers about the measures taken on their services to prevent banned groups/channels/communities from being recreated.

Provider	Measures in place to prevent banned groups/channels/communities from being recreated
Google YouTube	• Automated and machine learning systems • Multiple indicators used
Meta Facebook Instagram	• Strategic disruption of networks targeted at banned group's presence on Meta's services • Identifying signals that indicate a banned organisations presence • Ongoing enforcement sweeps against bad actors • Automatically disabling pages/groups with names associated with certain Dangerous Organisations and Individuals.
Reddit	• Subreddit ban evasion detection tooling • Several indicators used
WhatsApp	• Corresponding ban of admin(s)
Telegram	• Removing owners and administrators of infringing Communities • Minimal number of indicators

Recommender systems

Risks

Recommender systems determine what will be promoted to a service user based on many factors. Machine learning techniques are often used to identify user attributes and patterns and make recommendations to achieve particular goals, based on a range of data and signals on the service. There are many positive outcomes from recommender systems. For example, recommender algorithms that prioritise time spent reading or reacting to a post and then serve up similar content in the future can result in people seeing things they find interesting, entertaining or valuable.[99] However, there are risks if the objective of a recommender system is to deliver greater engagement without regard to safety. Recommender systems that prioritise maximising engagement run the risk of exploiting people's biases and drawing them to shocking and extreme content.

Recommender systems have been criticised for facilitating online radicalisation by progressively serving increasingly extremist and inflammatory material to maximise engagement. For some individuals, continuous exposure to TVE and other forms of hateful propaganda can have serious adverse effects by normalising prejudice and hatred and encouraging them to hold terrorist or violent extremist attitudes. Investigations into the motivations behind the Christchurch and Buffalo mass shootings have emphasised that both perpetrators were racially motivated violent extremists who were largely radicalised and inspired by extremist content and communities they discovered online.[100]

Without appropriate safeguards, recommender systems can support the aim of bad actors who deliberately seek to spread TVE online to glorify the actions of terrorists and violent extremists, promote their hateful ideologies, undermine social cohesion, and jeopardise public safety by inspiring copy-cat attacks.

In addition, algorithmic amplification of TVE – such as the recirculation of footage from livestreamed terror attacks – can inflict further pain and trauma on victims and their loved ones and distress members of the broader public.

---

[99] eSafety Commissioner, 'Recommender systems and algorithms – position statement', as updated 8 December 2022, accessed 4 June 2024, URL: https://www.esafety.gov.au/industry/tech-trends-and-challenges/recommender-systems-and-algorithms

[100] Royal Commission of Inquiry into the Terrorist Attack on Christchurch Mosques on 15 March 2019, 'Report: Royal Commission of Inquiry into the terrorist attack on Christchurch masjidain on 15 March 2019', 26 November 2020, accessed 4 June 2024, URL: https://christchurchattack.royalcommission.nz/; Office of the New York State Attorney General, 'Investigative Report on the role of online platforms in the tragic mass shooting in Buffalo on May 14, 2022', 18 October 2022, accessed 31 January 2024, URL: https://ag.ny.gov/sites/default/files/buffaloshooting-onlineplatformsreport.pdf

eSafety asked providers to report on measures taken to safeguard against the amplification of TVE-related harms by recommender systems.

Safeguarding recommender systems against TVE-related harms

There are a range of measures available to service providers to safeguard recommender systems against contributing towards TVE-related harms on their services. Providers can test and update recommender systems to reduce the risk that TVE is amplified. This process may involve initiatives such as internal audits, external audits, risk and impact assessments, and a/b testing.[101] Recommender systems can be programmed to stage positive interventions in circumstances where a service identifies that a user is actively engaging with, or searching for, TVE material. For example, it can promote deradicalising content in the feeds of that user or serve targeted pop-up notifications to counter any terrorist or violent extremist narratives.[102] This technique is also known as 'off-ramping'.

eSafety asked service providers about the tools, policies and processes they have in place to safeguard against the amplification of TVE-related harms by recommender systems, in relation to sections 6 and 11 of the Determination.

Table 20: In response to the notices, the following information was given by providers about the measures in place to prevent amplification of TVE via recommender systems.

Provider	Recommender systems tested to prevent amplification of TVE	Interventions to prevent amplification of TVE	Detail of interventions to prevent amplification of TVE
Google YouTube	Yes	Yes	• Removing violative content • Age-restrictions for content inappropriate for under 18 year olds • Training systems to elevate authoritative sources (eg regarding breaking news, politics, media and scientific information) higher in search results • Promoting authoritative sources in search results and in the event of 'breaking news' • Rewarding trusted creators through YouTube Partner Program

[101] A method used to compare two versions of something, such as a service or service feature, to determine which one performs better against predetermined criteria.

[102] Global Internet Forum for Countering Terrorism, 'GIFCT Technical Approaches Working Group: Gaps analysis and recommendations for deploying technical solutions to tackle the terrorist use of the internet', July 2021, accessed 4 June 2024, URL: https://gifct.org/gifct-resources-and-publications/

			• Providing information panels on videos and searches related to topics 'prone to misinformation'
Meta Facebook Instagram	No[103]	Yes	• In answer to a question about whether Meta had interventions in place to prevent the amplification of TVE via its recommender algorithms on Facebook and Instagram, Meta referred to the information it provided regarding the measures it takes to remove TVE from its services
Reddit	Yes	Yes	• Reddit periodically rates communities based on the content within those communities using an internal taxonomy rating system • Communities must meet certain size and activity thresholds to be eligible for rating, and content from unrated communities is not eligible for recommendation • Content must achieve a suitability score to be eligible for recommendation surfaces, such as home feed suggestions • Reddit's subreddit structure limits virality.

[103] Meta reported that it had not undertaken testing of its recommender system during the report period to ensure it did not amplify TVE.

6. Transparency summaries: Individual provider responses

Unless otherwise specified, the information contained in this report, and summarised in the following individual service provider responses, pertains to the report period (1 April 2023 to 29 February 2024). The tools, policies and processes that were in effect during the report period may have changed since.

Google summary

Overview

Google LLC. was asked about three services it provides: YouTube, Drive, and Gemini.

1. Questions about Google's definitions of 'terrorist material and activity' and 'violent extremist material and activity'

A. YouTube and Drive

In response to questions about how YouTube and Drive define 'terrorist material and activity' and 'violent extremist material and activity' or different but equivalent terms for the purposes of their terms of service and community guidelines, Google stated that YouTube and Drive do not use the term 'terrorist material and activity' and instead use the broader term 'violent extremist content.'

For the purposes of YouTube and Drive, Google defined 'violent extremist content' as

> content produced by or in support of terrorists and other violent organisations and movements that pose real-world harm. This includes (but is not limited to) content used to recruit for terrorist organisations, incite violence, glorify terrorist attacks, and promote acts of terrorism.

Google stated that YouTube's Community Guidelines prohibit 'violent extremist content' under its 'violent extremist and criminal organisations policy'[104], in addition to policies against hateful[105], violent or graphic content[106], and that 'terrorist material and activity' is prohibited under Drive's 'Violent Organisations and Movements Policy[107]' as well as policies against 'Hate Speech' and 'Violence and Gore'. Drive's 'Violent Organisations and Movements Policy' prohibits 'known violent non-state organisations and movements from using Drive for any purpose', and that it determines such organisations and movements through 'a variety of factors and inputs, including, but not limited to designated terrorist groups compiled by democratically elected governments such as the U.S Government and the U.N.'

Google also stated that YouTube will terminate any channel where it has 'reasonable belief that the account holder is a member of a designated terrorist organisation, including organisations identified by the United Nations'.

Google stated that for both Drive and YouTube, an educational, documentary, scientific or artistic (**EDSA**) exemption may apply to permit content related to violent non-state organisations or terrorist organisations that is shared for an EDSA purpose.

B. Gemini

Google stated that its 'Generative AI Prohibited Use Policy'[108]

> prohibits performing or facilitating dangerous, illegal or malicious activities, including promoting or generating violent extremism or terrorist content. These concepts are broadly defined to include content that relates to, incites or celebrates terrorism or violent extremism.

Google stated that it 'considers a number of factors and inputs' to determine what is violent extremist and terrorist content, 'including but not limited to terrorist groups compiled by democratically elected governments such as the U.S Government and the U.N.'

---

[104] Google, 'Violent extremist or criminal organizations policy', URL: https://support.google.com/youtube/answer/9229472?hl=en. URL supplied by Google on 22 May 2024.
[105] Google, 'Hate speech policy', URL: https://support.google.com/youtube/answer/2801939?ref_topic=2803176, URL supplied by Google on 22 May 2024.
[106] Google, 'Violent or graphic content policies', URL: https://support.google.com/youtube/answer/2802008?ref_topic=2803176, URL supplied by Google on 22 May 2024.
[107] Google, 'Violent organisations and movements policy', URL: https://support.google.com/docs/answer/148505?visit_id=638471172322536889-133621207&hl=en&rd=1#zippy=%2Cdangerous-and-illegal-activities%2Cviolent-organizations-and-movements%2Cviolence-and-gore%2Chate-speech. URL supplied by Google on 22 May 2024.
[108] Google, 'Generative AI prohibited use policy', 14 March 2023, URL: https://policies.google.com/terms/generative-ai/use-policy. URL supplied by Google 22 May 2024.

2. Thresholds/criteria to determine action on TVE breaches

Google was asked if it had criteria or thresholds in place to determine what action would be taken when TVE was identified on YouTube, Drive, and Gemini. Google provided the following information:

A. YouTube

Table A

Actions taken on accounts or content when TVE was identified	Criteria/thresholds reported for YouTube
Permanent account ban	Google stated that YouTube will 'terminate a channel' when: • The channel is dedicated to policy violating content • The channel has received three strikes in a 90 day period • There is a single instance of an egregious policy violation, e.g., the channel is connected to a known terrorist organisation. • There is otherwise a violation of YouTube's Terms of Service.
Temporary suspension	Google stated that YouTube does not apply temporary suspensions, but users that receive account strikes may be temporarily restricted from accessing 'certain features or functions for a period of time, such as posting content.'
Account strikes	Google stated that YouTube uses an account strike policy to take graduated enforcement actions against accounts that breach its rules without meeting the threshold for an immediate and permanent ban. It said this graduated enforcement consists of: • Warning: Typically applied in response to a user's first violation. Users must take policy training to have the warning expire after 90 days. However, if the user violates the same policy within 90 days, they will receive their first strike. • First strike: The user's public content is set to private and they are restricted from using various features including uploading videos or starting livestreams for a period of 1 week. • Second strike: Applied if the user receives a second strike within the same 90-day period as the first strike. Prevents the user from posting content for 2 weeks. • Third strike: If a user receives 3 strikes in the same 90-day period, their channel will be permanently removed from YouTube.
De-prioritisation in recommender system	Google stated that 'content that is violative of YouTube's Community Guidelines is removed and not recommended by

	the YouTube recommender system', and content that 'comes close to, but does not breach' YouTube's Community Guidelines may not be recommended in YouTube's recommender system.
Limiting reach	Google stated that graphic or violent content permissible under an EDSA exemption may be placed under an age-restriction, making it unavailable to users under 18 years, or non-logged in users, if it is not suitable for all audiences.

B. Drive

Table B

Actions taken on accounts or content when TVE was identified	Criteria/thresholds reported for Drive
Permanent account ban	Google stated that it 'will terminate a Google Account connected to a user's Drive Account if the account is confirmed to be owned or operated by a known terrorist or violent organisation.'
Limiting reach	Google stated that 'The ability to share (or a third party's ability to access) content violative of Drive's policies will be disabled.'

eSafety notes that limiting bans to accounts on Drive that are 'owned or operated by a known terrorist or violent organisation' may mean that terrorists and violent extremists who are not associated with a specific organisation – such as the Christchurch attacker – may not be banned.[109]

C. Gemini

Table C

Actions taken on accounts or content when TVE was identified	Criteria/thresholds reported for Gemini
Permanent account ban	Google stated that it 'may terminate a Google Account where the user has materially or repeatedly breached Gemini's Terms of Service (including the Generative AI Prohibited Use Policy)'.

[109] The Christchurch attack led to a system, set up by the GIFCT and of which Google is a member, for dealing with material that is not associated with a specific terrorist group

3. Questions about reporting of TVE

A. In-service reporting of TVE on Google services

In response to questions about whether users could report instances of TVE to Google within its services (as opposed to navigating to a separate webform or email address), Google responded:

Table D

Service	In-service reporting option Yes/No	Reporting categories
YouTube	Yes	• Promotes Terrorism • Hateful or abusive content • Violent or repulsive content
Drive (Consumer version; content when it is shared)	Yes	• Violent organisations and movements content • Violence • Hate speech

Google stated that for both YouTube and Drive, Google may review flagged content for violations of all abuse categories regardless of the option selected by the user to report the content.

B. Reporting mechanisms for other entities to report TVE

i. YouTube

In answer to a question about having separate reporting mechanisms for other entities to report TVE, Google responded that YouTube does have reporting mechanisms (separate from users in general) for law enforcement, Trusted Flaggers, and regulatory or other public authorities.

Google stated that YouTube operates a 'Priority Flagger Program' which is available to 'Government agencies, civil society groups, and NGOs with an identified expertise in recognising and fighting harm online in at least one policy area'. Google stated that participants in its 'Priority Flagger Program':

- Receive training in enforcing YouTube's Community Guidelines
- Are given priority review when they flag suspected violative content
- Have a direct line of communication with YouTube's Trust and Safety teams

Google drew a distinction between its voluntary 'Priority Flagger Program' and legally required content reporting channels such as the EU Digital Services Act (**DSA**) Trusted Flagger Requirement.

ii. Drive

Google stated that Drive does not have a separate or dedicated 'Priority Flagger Program' for TVE material. However, Google stated that it is 'growing the program to other products and policies where Google is seeing demand', and that the initial focus for Drive is introducing priority flaggers for Child Safety, Scams, and Misinformation.

iii. Gemini

Google stated that Gemini does not have a separate or dedicated 'priority flagger program' for TVE material.

4. Questions about proactive detection

A. Detecting known material using hash matching

i. Known TVE images

In response to questions about hash matching for known TVE images, Google provided the following information:

Table E

Parts of service	Used image hash matching tools	Names of tools used
YouTube		
YouTube	Yes	MD5/SHA256
YouTube profile picture		
YouTube video thumbnails		
Drive		
Drive (Consumer version; stored content)	No	
Drive (Consumer version; content when it is shared)	Yes	MD5/SHA256

eSafety notes that the tools used by Google are cryptographic hashing tools, which only detect exact matches, rather than perceptual hashing tools (such as PhotoDNA) that can also detect variations of material. Detection of variations is important for preventing the dissemination of material, particularly in circumstances where material has the potential to be edited and go viral. For example, following the Christchurch attack Facebook identified 800 visually distinct versions of the attack video within the first days.[110]

In response to why hash matching tools are not used on stored content on Drive, Google stated that

> Google Drive is predominantly a file storage service only accessible by the account owner. End-users use Drive for a variety of reasons, including that it is secure. End-users therefore have a reasonable expectation of privacy.

It added that it draws a distinction between privately stored content and content that users make publicly accessible and that

> Different considerations arise where that end-user has elected to share or disseminate a file more broadly and the real risk of harm that may actuate to other end-users as a result of such sharing or dissemination.

Google also stated that while hash-matching is effective at detecting specific images, it is less effective at determining context or purpose which is often essential for assessing whether TVE-related material is harmful and/or illegal, or if it is being used for legitimate, non-malicious purposes such as academic research or journalism. Google stated that

> of those parts of the service that Google does deploy automated tools, 96% of all unique items flagged require human review. In Google's experience, even with the use of human reviewers, it may still not always be possible to accurately determine either the context or the intent for why a certain piece of content may be stored within a personal private file storge (rather than sharing where ascertaining context is clearer).

Google further stated that unlike child sexual abuse material, there is no generally agreed or uniform definition of TVE, and that there is a risk that

> Errors in enforcement (or in accurately detecting illegal TVE material) may cause other significant harms, including adversely affecting the rights of users to privacy, freedom of expression and access and use of information for legitimate and lawful purposes.

[110] A Further Update on New Zealand Terrorist Attack | Meta (fb.com), accessed 22 July 2024, URL: https://about.fb.com/news/2019/03/technical-update-on-new-zealand/

Google stated that it relies on user reporting and 'Engagement of external partners to detect potential violative drive files that contain TVE "off-platform"' to otherwise detect known TVE images in content stored on Drive.

ii. Known TVE video

In response to questions about hash matching for known TVE video, Google provided the following information:

Table F

Parts of service	Used video hash matching tools	Names of tools used
YouTube		
YouTube	Yes	MD5/SHA256
Drive		
Drive (Consumer version; stored content)	No	
Drive (Consumer version; content when it is shared)	Yes	MD5/SHA256

In response to why hash matching tools are not used to detect known TVE videos stored in Drive, Google referred to its reasons for not using such tools to detect images stored in Drive, including a higher expectation of privacy and challenges using tools without the availability of context.

iii. Known TVE written material

In response to questions about hash matching for known TVE written material on Drive, such as manifestos or text promoting, inciting, instructing TVE, Google provided the following information:

Table G

Parts of service	Used hash matching tools for written material	Names of tools used
Drive		
Drive (Consumer version; stored content)	No	
Drive (Consumer version; content when it is shared)	Yes	MD5/SHA256

In response to why hash matching tools are not used to detect known TVE written material stored in Drive, Google referred to its reasons for not using such tools to detect images stored in Drive, including a higher expectation of privacy and challenges using tools without the availability of context.

iv. Sources of TVE hashes

Google reported that it sourced its hashes of known TVE images, video, and written material from the following databases:

- The GIFCT's Hash-Sharing Database
- Google's internal hash database

Google stated that its internal hash database contains hashes of content that have been detected on Google services and hashes that Google has accepted from the GIFCT. Google stated it ingests all hashes from the GIFCT's database and then will 'undertake a review to verify whether content that matches those hashes violates Google's policies.'

Google also stated that its internal hash database is available for 'all Google services or products that use hash-matching to detect TVE.'

B. Detecting new TVE material

i. New or 'unknown' images

In response to questions about the detection of new (or 'previously unknown') TVE images, Google provided the following information:

Table H

Parts of service	Used tools for images	Names of tools used
YouTube		
YouTube profile picture	Yes	Proprietary Google image detection technology
YouTube video thumbnails		
Drive		
Drive (Consumer version; stored content)	No	
Drive (Consumer version; content when it is shared)		

In response to why automated tools are not used to detect new TVE images on any part of Drive, Google referred to its reasons for not using hash matching tools to detect known TVE stored in Drive.

ii. New or 'unknown' TVE videos

In response to questions about the detection of new (or 'previously unknown') TVE videos, Google provided the following information:

Table I

Parts of service	Used tools for videos	Names of tools used	Whether tools are video and/or audio classifiers, or others
YouTube			
YouTube	Yes	• Proprietary Google classifier technology A	Video, audio, and text
Drive			
Drive (Consumer version; stored content)	No		
Drive (Consumer version; content when it is shared)	Yes	• Proprietary Google classifier technology A • Proprietary Google hashing technology	Video

In response to why automated tools are not used to detect new TVE videos in stored content on Drive, Google referred to its reasons for not using hash matching tools to detect known TVE material stored in Drive, including a higher expectation of privacy and challenges using tools without the availability of context.

iii. Text Analysis to detect TVE

In response to questions about technology used to detect phrases, codes, hashtags indicating likely TVE in text (for example manifestos or text promoting, inciting, instructing TVE), Google provided the following information:

Table J

Parts of service	Used text analysis tools	Names of tools used
YouTube		
YouTube username	Yes	BERT (Bidirectional Encoder Representations from Transformer)
YouTube account description		
YouTube video titles		
YouTube video descriptions		

YouTube comments sections		
YouTube playlist titles	No	
Drive		
Drive (Consumer version; stored content)	No	
Drive (Consumer version; content when it is shared)		
Drive filename	Yes*	*Google clarified that there is no ongoing monitoring or scanning, but Google will scan for duplicates of known violative files on 'an ad-hoc or case by case basis'.

In response to why it does not use any technology to scan YouTube playlist titles for indications of likely TVE, Google stated that playlists are lists of videos on YouTube which are already available on YouTube and subject to its Community Guidelines. Google stated that 'if the content within the playlist itself is not violative', the presence of a particular keyword in a playlist title is 'unlikely to indicate violative conduct or behaviour'.

In response to why it does not use any technology to scan content on Google Drive for indications of likely TVE in text, Google stated

> While using keywords, hashtags or codes can be an effective method of detecting potential TVE activity on certain services, such as in comments posted on social media, in Google's experience it has not proven an effective tool to detect potential TVE in files or documents that are of the nature typically contained in a user's Drive for both stored or shared content.

iv. Source of phrases, codes, hashtags

Google stated that YouTube sources phrases, codes, and hashtags likely to indicate TVE from:

- Google and YouTube Trust and Safety teams
- External partners who provide Google's Trust and Safety teams with insights on potential risks in TVE and other areas
- Human review of content flagged and confirmed as TVE on YouTube
- GIFCT threat analysis briefings

C. Languages covered by language analysis tools

In response to questions about the languages covered by Google's language analysis tools, Google stated that its tools for detecting new TVE videos (including livestreams) and phrases, codes, and hashtags indicating likely TVE are capable of operating in the following languages:

Table K

Afrikaans	Albanian	Arabic	Aragonese	Armenian	Asturian
Azerbaijani	Bashkir	Basque	Bavarian	Belarusian	Bengali
Bishnupriya Manipuri	Bosnian	Breton	Bulgarian	Burmese	Catalan
Cebuano	Chechen	Chinese (Simplified)	Chinese (Traditional)	Chuvash	Croatian
Czech	Danish	Dutch	English	Estonian	Finnish
French	Galician	Georgian	German	Greek	Gujarati
Haitian	Hebrew	Hindi	Hungarian	Icelandic	Ido
Indonesian	Irish	Italian	Japanese	Javanese	Kannada
Kazakh	Kirghiz	Korean	Latin	Latvian	Lithuanian
Lombard	Low Saxon	Luxembourgish	Macedonian	Malagasy	Malay
Malayalam	Marathi	Minangkabau	Mongolian	Nepali	Newar
Norwegian (Bokmal)	Norwegian (Nynorsk)	Occitan	Persian (Farsi)	Piedmontese	Polish
Portuguese	Punjabi	Romanian	Russian	Scots	Serbian
Serbo-Croatian	Sicilian	Slovak	Slovenian	South Azerbaijani	Spanish
Sundanese	Swahili	Swedish	Tagalog	Tajik	Tamil
Tatar	Telugu	Thai	Turkish	Ukrainian	Urdu
Uzbek	Vietnamese	Volapük	Waray-Waray	Welsh	West Frisian
Western Punjabi	Yoruba				

D. Action taken on TVE

In response to questions about what action was taken when known and new TVE images, video, and known written material were detected by its tools, Google provided the following information:

i. YouTube

- The TVE content is removed from the service and an email is sent to the end-user advising them of this action.

- New TVE content will be labelled and used to detect and remove re-uploads of the same content on YouTube. It may be shared with other Google services to aid detection.
- Actions taken against the user may include warnings, strikes, and/or termination of the user's YouTube channel depending 'on the severity of the violation.' YouTube will terminate any YouTube channel where it has a 'reasonable belief that the account holder is a member of a designated terrorist organisation' (such as those identified by the United Nations or the U.S.).
- Google may escalate and report to law enforcement where Google believes there is a credible threat to life or serious harm.

ii. Drive

- Google disables the ability of the owner to share the material to other users and the content will be rendered inaccessible to third parties.
- New TVE content will be labelled and used to detect and remove re-uploads of the same content on Drive. It may be shared with other Google services to aid detection.
- Google will disable a Google Account where it has reasonable belief that the account holder is a member of a designated terrorist organisation.
- Google may escalate and report to law enforcement where Google believes there is a credible threat to life or serious harm.
- Google may undertake a broader review of other content linked to the user's Drive to identify any further or additional shared TVE material.

iii. Action taken on likely written TVE

Google stated that violative content in YouTube comments is automatically removed when it is detected by YouTube's automated flagging systems. Google stated that videos and channels – including username, account description, video titles and descriptions – are reviewed by human moderators to confirm that they are violative. Google also referred to the steps it takes when it detects known TVE images and videos and written TVE.

When asked if Google blocks words or phrases that it detects indicating likely TVE to users searching for them, Google responded that '[w]hile YouTube does not prevent a user entering a particular search term, YouTube's systems are designed to prioritise relevance, quality and engagement on YouTube Search… YouTube Search raises authoritative sources (for example credible news sources on violent extremist or terror events) and reduces borderline content, including those related to TVE which comes close to, but does not quite violate our Community Guidelines, from being widely viewed'.

E. Livestreamed TVE

i. Detecting livestreamed TVE on YouTube

In response to questions about the measures YouTube had in place to detect the livestreaming of TVE on YouTube, Google provided the following information:

Table L

Parts of service	Measures in place to detect TVE in livestreams?	Interventions used	Names of tools used
Livestream video	Yes	• Text classifiers • Video classifiers • Audio classifiers • Keyword detection	Proprietary Google Classifier Technology B
Live chat			
Text associated with livestream (title and description)			

Google stated that its livestreaming detection classifiers operate in the same languages as the tools it used to detect new TVE in videos and in written text (see **Table K**).

ii. Reducing the likelihood of TVE in livestreams on YouTube

In response to questions about the steps YouTube takes to reduce the likelihood that TVE could occur in livestreams, Google stated that it used the following measures:

- Priority reviews of reports about livestreamed content
- Restrictions for those who have previously violated TOS or community guidelines/standards. Users must not have any live-streaming restrictions on their account in the last 90 days (i.e. a prior strike).Minimum audience requirements
- Requirements that users verify their Account or Channel by phone number to enable livestreaming
- A 24-hour waiting period before a user can deploy the livestream functionality after enabling it on their account
- Additional restrictions for livestreams from mobile devices. Users must have at least 50 subscribers, and users with less than 1000 subscribers may have the number of viewers limited by YouTube. All archived live-streams are set to private by default.

- Prohibitions on livestreams that 'show someone holding, handling, or transporting a firearm'. Channels that violate YouTube's firearms policy may lose their ability to livestream.

iii. In-service reporting of livestreams by users that are not logged in to YouTube/Google

In response to a question, Google stated that there is no mechanism to enable users that are not logged in to YouTube/Google to make an in-service report about livestreamed TVE.

In response to a question about the alternative steps Google takes to ensure that its reporting mechanisms for livestreamed TVE are clear and readily identifiable (as expected by sections 13 and 15 of the Determination), Google stated

> YouTube provides all users with clear and readily available information on how to report videos, channels and other content, and check on the outcome of a report here [link][111]. Users can also make a "legal report" either in-service or via an external webform, available here [link][112].
>
> While YouTube supports in-service functionality that enables users to flag potentially violative or illegal content, in YouTube's experience user flagging is nonetheless highly susceptible to abuse and manipulation – for example, users flagging content for non-legitimate or malicious reasons.

Google added that during the 6-month period ending September 2021

> less than 2% of the more than 32 million videos flagged globally for review under YouTube's Community Guidelines were ultimately removed after human review of that content.

F. Blocking links to TVE material

i. Detection and sources of URLs

Google was asked about its use of lists or databases to proactively detect and block URLs linking to TVE on other platforms. Specifically, Google was asked about:

- Known URLs linking to websites/services operated by individuals/organisations dedicated to the creation, promotion, or dissemination of TVE or other TVE-related activities
- URLs linking to known TVE material on other services/websites (which may not be dedicated to TVE)

---

111 YouTube, 'Report inappropriate videos, channels, and other content on YouTube', accessed 4 July 2024, URL: https://support.google.com/youtube/answer/2802027?hl=en&co=GENIE.Platform%3DAndroid#zippy=%2Creport-a-video%2Creport-a-short%2Creport-a-channel%2Creport-a-playlist%2Creport-a-thumbnail. URL supplied by Google.
112 Google, 'Report content for legal reasons', accessed 4 July 2024, URL: https://support.google.com/legal/answer/3110420?hl=en. URL supplied by Google.

- Join-links to groups, channels, communities, or forums on other services that were known to be associated with TVE.

Table M

Parts of service	Blocked URLs to websites/services dedicated to TVE	Blocked URLs linking to known TVE material on other services/websites	Blocked join-links to groups/channels on other services known to be associated with TVE	URL sources
YouTube account descriptions	Yes	Yes	Yes	Human review of suspicious video, channels and URLs. Google stated that it does not source URLs from external sources or lists.
YouTube video descriptions	Yes	Yes	Yes	
YouTube comments sections	Yes	Yes	Yes	

While Google does not source URLs from external sources, eSafety notes that Google and YouTube are members of the GIFCT, which makes hashes of URLs to known TVE available to its members.

ii. Action taken on accounts attempting to share blocked URLs/join-links

In response to questions about what action was taken when an account was detected attempting to share a blocked URL dedicated to TVE, a blocked URL linking to TVE on another website/service or a blocked join-link to groups channels on other services known to be associated with TVE, Google stated that:

> If YouTube detects URLs that are confirmed to link to TVE in violation of YouTube Policies, YouTube will remove the content displaying the URL... where new URLs that link to TVE content are confirmed, these may be added to the internal YouTube blocklist.

In addition to this response Google also referenced the steps it reported taking when it detects known TVE images, videos and written material and new TVE images and video.

G. Off-platform monitoring

In response to a question about whether Google used off-platform monitoring,[113] either provided internally or by third-party services, to identify accounts or channels dedicated to TVE on YouTube and Drive, Google stated that it has an internal specialist 'Trust and Safety Intel Team' that surveys breaking news developments and the wider internet to identify abuse trends 'which includes but is not limited to violent extremist and terrorist activity'. Google also said that it uses third-party vendors to 'provide additional expertise, resources, or to augment any gaps in coverage.'

Google stated that it takes into account government and other expert advice on violent extremist and terrorist threats, such as UN designations of terrorist organisations. Google further stated that it is a member of the GIFCT and the Christchurch Call, two organisations that provide an important role in:

> (a) sharing intel, best practices or threat analysis to online service providers; (b) participation in, and access to the GIFCT hash database; and (c) participation in the GIFCT Content Incident Protocol (CIP) that enables GIFCT member companies to quickly become aware of, have access to, and address harmful online content resulting from a terrorist or violent extremist event.

Google stated that YouTube has its own 'YouTube Intelligence Desk' which specialises in 'identifying new potential violative trends, including off-platform', and that Drive also engages third party agencies to detect Drive links to TVE material or activity being shared on non-Google platforms.

i. Off-platform monitoring by third-party services

In response to a question seeking a list of the third-party services that Google engages to perform off-platform monitoring for TVE-related threats, Google provided information about a third-party service it engaged to perform such work during the report period. Google noted that these third-party services or platforms may vary from time to time.

H. Percentage of reports sent for human review

In response to questions about the percentage of TVE reports sent for human review and the criteria and thresholds used to determine when reports are sent for review, Google provided the following information:

[113] Monitoring of activity on other services.

Table N

	Percentage of user reports of TVE sent for human review	Criteria and thresholds used to determine when a user report is sent for human review	Percentage of TVE detected through automated tools sent for human review	Criteria and thresholds used to determine when a report of TVE detected through automated tools is sent for human review
YouTube	99%*	Likelihood that the content is violative. Human review may not be required if there is 'high confidence' that the content is violative. For example, if the item is an exact match or duplicate of content that was previously assessed as violative by a human reviewer.	86.4%**	Likelihood that the content is violative. Human review may not be required if there is 'high confidence' that the content is violative. For example, if the item is an exact match or duplicate of content that was previously assessed as violative by a human reviewer.
Drive	100%	N/A	96%***	

*** Google reported that the 99% refers to videos uploaded from Australia that were sent for human review after first being flagged by a user or Priority Flagger and were subsequently confirmed to violate YouTube's policies for 'violent extremism and criminal organisations'.**

**** Google reported that the 86.4% refers to videos uploaded from Australia that were sent to human review after first being detected by automated flagging and were subsequently confirmed to violate YouTube's 'violent extremism and criminal organisations' policies.**

***** Google reported that the 96% refers to unique items that were sent for human review after being flagged by Drive's automated tools and were subsequently confirmed to violate Drive's policies for terror and violent extremism.**

I. Percentage of TVE detected proactively

Google was asked what percentage of TVE was detected proactively, compared to TVE reported by users, trusted flaggers, or through other channels for the following services:

Table O

Service	Percentage of TVE detected proactively	Percentage of TVE reported by users, trusted flaggers, other
YouTube	95.3%*	0.8% Priority Flaggers* 3.9% users*
Drive (consumer version)	~66%**	34%**

*** Google stated these figures represent the percentage of videos uploaded from Australia that violated YouTube's 'violent extremism and criminal organisations' policy that were first flagged by YouTube's automated detection tools or by users and Priority Flaggers.**

****Google stated that due to its data retention policies, some of the data requested by eSafety was no longer available and that these figures were calculations based on 'good-faith efforts and the 'best data that is currently available for the Reporting Period'.**

J. Appeals against TVE-related moderation

In response to a question about how many appeals were made by users for accounts banned or content removed for TVE, where the service was alerted by automated tools or user reports, and how many of those were successful, Google provided the following information:

Table P

How Google was alerted to TVE	Number of appeals made for accounts banned for TVE breach	Number of appeals that were successful for accounts banned	Number of appeals made for material removed for TVE breach	Number of appeals that were successful for material removed
YouTube				
Automated tools	0	0	251*	17*
User reports	0	0	20*	3*
Drive (consumer version)				
Automated tools	0	0	18**	1**
User reports	0	0		

*** Google stated that these figures are Australian only.**

**** Google reported that due to its data retention policies, it did not have the data necessary to distinguish appeals and reinstatements based on whether the material had been detected via automated tools or from a user report. Google noted that the figures provided are global.**

5. Questions about resources, expertise, and human moderation

A. Trust and Safety

i. Trust and Safety and other staff

eSafety referred in the Notice to the fact that, in January 2023, Google had announced reductions to its staffing numbers.[114] Google was asked to provide the number of staff that were employed or contracted by Google to carry out certain functions at both the beginning and the end of the report period. Google provided the following information:

Table Q

Category of staff	1 April 2023	29 February 2024
Engineers employed by Google focused on trust and safety	1305	1294
Content moderators employed by Google	316	341
Content moderators contracted by Google	39,606	39,552
Trust and safety staff employed (other than engineers and content moderators)	1,416	1,265

Google noted that it has 'numerous teams that mitigate systemic risk and operate both vertically for a particular product area, and also horizontally across Google services, including Drive, YouTube and Gemini.'

Google also stated that to the extent there were variations in headcount between the two dates requested by the Notice, 'these should not be assumed as necessarily the result of the announcement in January 2023 with regard to an overall reduction in staffing numbers across Alphabet'.

ii. Trust and safety dedicated to minimising TVE

Google was asked if it had a dedicated trust and safety team(s) responsible for minimising TVE on YouTube. Google stated that YouTube does not have 'a set team responsible for minimising

[114] Google, 'A difficult decision to set us up for the future', 20 Jan 2023, accessed 29 January 2024, URL: https://blog.google/inside-google/message-ceo/january-update/

TVE only.' Following subsequent correspondence with Google, Google provided the following information about the number of humans that were employed to evaluate YouTube content in English (which it said 'would include Australia and TVE content), and the number of humans that were employed to conduct 'language agnostic reviews':

Table R

Category of content reviewer	31 December 2023	31 March 2024[115]
English language reviewers	3,455	3,243
Language agnostic reviewers	9,813	9,322

Google stated that 'agnostic reviews are primarily done when no language is needed to conduct the review (e.g., adult content) or in specific cases when YouTube cannot identify the language.'

iii. Surge teams to respond to a TVE crisis

Google was asked if it had a surge team to respond to TVE crises, such as a livestreamed attack with content disseminated on YouTube. Google answered 'no'.

It added that YouTube has 'rapid response capabilities' to ensure that it responds to major incidents, including livestreamed terrorist attacks.

B. Languages human moderators operate across

In response to a question about the languages that its human moderators operate across (both employees and contractors), Google provided the following:

Table S

Languages covered by employees (all languages)	Languages covered by contractors (all languages)		
• English	• Afrikaans • Amharic • Arabic • Azerbaijani • Belarusian • Bengali • Bosnian • Bulgarian • Burmese • Cantonese	• Hindi • Hungarian • Igbo • Indian Languages • Indonesian • Irish • Italian • Japanese • Kazakh • Khmer	• Portuguese • Portuguese-BR • Punjabi • Romanian • Russian • Serbian • Sinhalese • Slovenian • Somali • Spanish

[115] Google stated that it was unable to provide employee data specific to the report period because it 'has standardised its processes to capture data at particular intervals'.

	• Croatian • Czech • Danish • Dutch • Dutch • English • Estonian • Ethiopian-Amharic • Ethiopian-Oromo • Ethiopian-Tigriniya • Filipino • Finnish • French • German • Greek • Gujarati • Hausa • Hebrew	• Korean • Kurdish • Laos • Latvian • Lithuanian • Macedonian • Malay • Malayalam • Mandarin • Mandarin/Cantonese • Marathi • Norwegian • Oriya • Oromo • Pashto • Persian • Polish	• Swahili • Swedish • Tagalog • Tajik • Tamil • Telugu • Thai • Tigrinya • Turkish • Ukrainian • Urdu • Uyghur • Uzbek • Vietnamese • Yoruba • Zulu

C. Median time to reach an outcome to a user report of TVE

Google was asked to provide the median time taken to reach an outcome[116] after receiving a user report about TVE for the following services:

Table T

Parts of the service	Reports from users globally	Reports from users in Australia
YouTube	15 minutes for automated review of the flag. Approximately 4.4 hours for flags referred to human review.*	Google reported that this information is not available.
Drive (Consumer version) (Content when shared)	10.2 hours**	2.9 hours**

*** Google reported that these two figures reflected that YouTube has two processes for reviewing a user flag. The first process involves automated review of the flag to determine whether it should be sent for**

[116] Defined in the Notice as a calculation from 'the time that a user report is made, to a content moderation outcome or decision, such as removing the content, banning the account, or deciding that no action should be taken.'

human review. The second process involves prioritising referred content for human review where a policy decision is then made on what action YouTube should take.

Google reported that YouTube's figures were based on data that is not TVE-specific and were from outside the report period. Google stated that YouTube did not have data to distinguish the median time to enforce user flags based on country of origin or specific to its TVE policies.

Following a request for clarification by eSafety, Google stated that the data is based on a study completed in July 2022 and that it relates to user flags on videos that are potentially violative of community guidelines, including guidelines related to TVE.

**** Google reported that these figures refer to the median time taken from when a user flag is first received to when an outcome is reached.**

Google stated that the time taken to reach an outcome on both Drive and YouTube depends on a range of signals or factors, including the likelihood that the content is violative or where there is a risk of serious harm.

6. Questions about steps to prevent recidivism

A. YouTube

i. Measures and indicators

In response to a question about the measures taken to prevent recidivism for TVE-related breaches on YouTube, Google stated that

> YouTube has processes in place to terminate accounts related to users who have been previously terminated because of violations by using relatedness signals to determine if two channels are related or not.

Google listed multiple indicators[117] that YouTube used to detect users who have previously been banned for TVE-related breaches. eSafety has chosen not to publish these indicators to prevent the information being misused.

Google stated that YouTube used all indicators by default in all instances where an account was banned to prevent recidivism by that user.

ii. Preventing banned TVE channels from being recreated

In response to a question about the measures YouTube took to prevent banned TVE channels from being recreated, Google stated that if a user's channel has been terminated or restricted

[117] eSafety uses the following terms to give an impression of the extent of the indicators used by services in the table below, rather than publishing the specific indicators which could be misused:
- Minimal: A small number
- Several: A moderate number
- Multiple: A significant number.

due to a strike, that user will be prohibited from 'using another channel to circumvent these restrictions.' Google also stated attempts to circumvent a prohibition may result in the immediate termination of the channel.

Google stated that YouTube uses various automated tools to detect the re-upload of violating content and that it also uses indicators to detect recidivist channels. Google said it also relied on user reporting to identify channels that may have been re-created to circumvent a ban.

iii. Applying TVE-related bans to associated accounts

Google was asked, when it took action against a user for a TVE-related breach, whether it applied bans to associated accounts. eSafety defined 'associated accounts' as 'other users who are associated with the banned user', such as accounts subscribed to the same TVE-related channels as the banned account. Google stated 'we are unclear how an "associated account" is defined in the context of this question', and responded by referring to the steps it takes to prevent users from re-creating new YouTube channels to circumvent bans.

B. Drive

i. Measures and indicators

In response to a question about the measures Google takes to prevent recidivism for TVE-related breaches on Drive, **Google listed a minimal number of indicators that it used to detect users that have previously been banned for TVE breaches.** eSafety has chosen not to publish these indicators to prevent the information being misused.

Google stated that Drive used all indicators by default in circumstances where an account was banned to prevent recidivism by that user.

C. Sharing of banned account details with other entities

Google was asked if it shared details of accounts banned for TVE with the following entities:

Table U

Entity	Shared details of accounts banned for TVE	Details provided by Google
Other service providers	No	Google referred to its Privacy Policy, stating that: 'Google does not share personal information with companies, organisations, or individuals outside Google, except in the following cases: (i) with the user's consent, (ii) with domain administrators, (iii) for external

		processing, if needed, (iv) for legal reasons (for example, to respond to any applicable law, regulation, or legal process).' Google noted that it participates in industry forums such as the GIFCT.
Law enforcement	Yes	Google stated that 'If Google reasonably believes that disclosing user information can prevent someone from dying or from suffering serious physical harm, Google may voluntarily provide this information to a government or law enforcement agency'. Google stated that this includes, for example, cases of bomb threats, school shootings, kidnappings.
Regulatory or other public authorities	Yes	Google stated that when 'Government agencies from around the world ask Google to disclose user information. Each request is carefully reviewed to ensure it satisfies applicable laws. The number and types of requests received are shared in the Transparency report.'[118]
Global Internet Forum to Counter Terrorism	No	'GIFCT does not have a process that enables the sharing of specific account details amongst participating members.'
Civil society groups	No	Referred to response regarding 'other service providers'.

7. Questions about recommender systems

A. Preventing amplification of TVE

i. Recommender algorithm – interventions

In answer to a question about whether YouTube had interventions in place to prevent the amplification of TVE via its recommender algorithm, Google provided the following measures:

- Removing content that violates YouTube's Community Guidelines
    - Google stated that if 'content is removed, that content cannot be amplified by YouTube's recommender systems.'
- Age-restricting content that may not be appropriate for users under 18 but has significant eEDSA value. This content will not be viewable to users that are below 18 years of age[119], or those that have not logged into an account.

[118] Google.com, 'Google Transparency Report', accessed 24 July 2024, URL: https://transparencyreport.google.com
[119] Users that have created a YouTube account with a registered age below 18 years old.

- Promoting authoritative sources by training YouTube's systems to elevate these sources higher in search results, particularly in contexts where accuracy and authoritativeness are important (e.g., breaking news, politics, media, and scientific information).
    - 'Content that comes close to, but does not cross the line of violating YouTube's Community Guidelines is not recommended to users or surfaced prominently in search results.'
    - YouTube uses human raters trained on 'public guidelines' to assess 'authoritative' or 'borderline' content. YouTube then feeds the 'consensus input' from its human evaluators into a 'well-tested machine learning system' to build models that 'help review hundreds of thousands of hours of videos every day in order to find and limit the spread of borderline content.'
- Rewarding trusted creators through the YouTube Partner Program (**YPP**) which enables users that meet eligibility thresholds to monetise their content through advertising.
- The eligibility thresholds relate 'to watch time and subscribers' but if a creator has activated ads monetisation for a video, and YouTube's reviewers and automated systems determine that it does not comply with YouTube's 'advertiser-friendly content guidelines'[120] then the video will have 'limited or no ads appear against it'[121].
    - In cases of severe or repeated violations of YouTube's monetisation policies, YouTube may suspend a creator's channel from the YPP.
- Showcasing 'high quality, authoritative news sources' that appear automatically for 'top' and 'breaking' news.
- Channels selected for this feature must follow the 'Google Search feature policies'[122] and 'Google News' content policies'[123] and Google reported that it uses various signals 'that may include channel quality and channel coverage of recent and relevant news events.'
- Providing 'information panels' on videos and searches 'related to topics that are prone to misinformation'.
    - These information panels 'show basic background info, sourced from independent, third-party partners, to give more context on a topic.'

[120] YouTube Help, 'Advertiser-friendly content guidelines', accessed 24 July 2024, URL: https://support.google.com/youtube/answer/6162278
[121] YouTube Help, '"Limited or no ads" explained', accessed 24 July 2024, URL: https://support.google.com/youtube/answer/9269824
[122] Google Search Help, 'Content policies for Google Search', accessed 24 July 2024, URL https://support.google.com/websearch/answer/10622781
[123] Publisher Centre Help, 'Google News Policies', accessed 24 July 2024, URL: https://support.google.com/news/publisher-center/answer/6204050?visit_id=637950

ii. Recommender algorithm – testing

In answer to a question about any testing YouTube performs to ensure that its recommender systems do not amplify TVE, Google provided the following measures:

- Violative view rate (**VVR**) – a metric that Google reported is key for assessing how quickly YouTube removes TVE material from its service (thereby preventing its amplification).
    - 'VVR is an estimate of the proportion of video views that violate our Community Guidelines in a given quarter (excluding spam). In order to calculate VVR, we take a sample of the views on YouTube and send the sampled videos for review. Once we receive the decisions from reviewers about which videos in the sample are violative, we aggregate these decisions in order to arrive at our estimate'.
    - 'As the overwhelming majority of violative content [that Google is becomes aware of] is detected by automated systems, YouTube's Violative View Rate (VVR) is a good indication of how well YouTube's automated systems are protecting the community. Although metrics like turnaround time to remove violative videos or number of takedowns are important, these statistics do not fully capture the actual importance of violative content on viewers and the extent of dissemination.'
- Community Guidelines development - Google reported that YouTube engages in ongoing reviews of its Community Guidelines to address new and emerging threats. Google added that this includes working with NGOs, academics, and other relevant experts, insights from the YouTube Intelligence Desk (see section G), testing of enforcement guidelines by content moderators, and 'regular meetings by YouTube's trust and safety specialists across the globe' to discuss enforcement of individual policies.
- YouTube Researcher Program – Academic researchers are allowed 'scaled access' to YouTube's data API for research projects.

Google also noted that since YouTube made changes to its Recommender System in 2019, many third-party independent studies have examined the effects or likely effects of this system on amplifying harmful content. Google stated that it considers these studies important for evaluating and tests of its recommender system. Google listed 6 studies and stated:[124]

[124] The studies listed by Google were:
- 'Examining the consumption of radical content on YouTube', Proceedings of the National Academy of Sciences (PNAS), 2021, URL: https://www.pnas.org/doi/10.1073/pnas.2101967118. URL supplied by Google.
- 'Algorithmic extremism: Examining YouTube's rabbit hole of radicalisation', The University of California, Berkley, the School of Information, 2020, URL: https://firstmonday.org/ojs/index.php/fm/article/view/10419. URL supplied by Google.
- 'A longitudinal analysis of YouTube's promotion of conspiracy videos', M., Faddoul et al., 2020, URL: https://arxiv.org/abs/2003.03318. URL supplied by Google.
- 'Social media, extremism, and radicalisation', Science Advances, 2023, URL: https://www.science.org/doi/10.1126/sciadv.adk2031. URL supplied by Google.

YouTube's observation of these studies is that they show that the issue of algorithmic radicalization is more nuanced than often portrayed in the media and there is no clear or consistent evidence that the recommender system has a significant impact on amplifying TVE content or radicalising individuals.

iii. Recommender algorithm – positive interventions

Google was asked if YouTube had systems in place to stage positive interventions, for example by promoting deradicalising content for at-risk users when a user sought out TVE material on the service. Google stated that it did have such measures in place and referred to the answers provided regarding measures to prevent the amplification of TVE material.

8. Questions about Generative AI safety

A. Labelling AI generated content

Google was asked if it took any steps to embed indicators of provenance – commonly known as 'watermarks' - into the material generated by its Gemini service to aid the proactive minimisation of unlawful and harmful material. Google provided the following information:

Table V

Type of content	Embedded indicators Yes/No	Perceptible marks[125] Yes/No	Tools used	Details of indicators
Images	Yes	No	SynthID for images	Gemini uses SynthID to embed digitally identifiable watermarks into the pixels of generated images. SynthID for images uses two deep learning models – one to apply and one to detect watermarks. The tool is designed to allow the watermark to 'remain detectable, even after modifications like adding filters, changing colours, and saving with various lossy compression schemes (commonly used for JPEG images).'
Text	Yes	No	SynthID for text	Gemini uses SynthID to add digitally identifiable watermarks into generated text. 'SynthID for text is designed to embed imperceptible watermarks directly into the

- 'Algorithmic recommendations have limited effects on polarisation: A naturalistic experiment on YouTube', N., Liu et al., 2023, URL: https://dcknox.github.io/files/LiuEtAl_AlgoRecsLimitedPolarizationYouTube.pdf?utm_source=pocket_saves&utm_medium=email. URL supplied by Google.

'YouTube recommendations point to more popular content – regardless of starting criterion', Pew Research Center, URL: https://www.pewresearch.org/internet/2018/11/07/many-turn-to-youtube-for-childrens-content-news-how-to-lessons/. URL supplied by Google.

[125] Perceptible being 'visible to the naked eye'.

				text generation process. It does this by introducing additional information in the token distribution at the point of generation... The final pattern of scores for both the model's word choices combined with the adjusted probability scores are considered the watermark.'

Google noted that although its embedded indicators are not perceptible to the naked eye, the SynthID tool can be used to scan an image to detect the digital watermark. Google stated that

> SynthID provides three confidence levels for interpreting the results for identification. If a digital watermark is detected, part of the image is likely generated using our AI models.
>
> Similarly, while there is no text watermark that is perceptible to the human eye, the pattern of scores can be compared with the expected pattern of scores for watermarked and unwatermarked text to detect if Gemini generated the text or if it might come from other sources

eSafety notes that SynthID is available to Vertex AI customers. In response to a supplemental question, Google reported that 'customers of Vertex are able to check content for their own watermarks' and that '[s]ome additional synthID features are available for users of Search'. Google provided the following link[126] for more information.

B. Preventing TVE and CSEA prompts

Google was asked about the measures it took to prevent end-users from entering prompts, or making requests, that might result in Gemini generating synthetic material that is likely to be illegal or seriously harmful. Google was asked to report on these measures as they related to the potential generation of TVE and CSEA material. Google stated that it uses 'prompt classifiers' to

> help determine whether or not a prompt is on a topic that could lead to an answer which violates our policies, and we leverage that to determine whether we should block the prompt altogether, or to dynamically re-prompt Gemini to generate a safer response. This includes prompts that seek child abuse material or may result in violent content.

Google also provided the following information in response to questions about user prompts:

[126] Google 'Get helpful context About this image', 10 May 2023, URL: https://blog.google/products/search/about-this-image-google-search/

Table W

Harm type	Banned keywords from text prompts?	Scanning user image prompts using hash matching and classifiers for TVE and CSEA	Any other interventions
CSEA	No[127]	Google stated that it 'uses both hash-matching (for user uploaded images) and machine learning classifiers (for text-based prompts) to identify potential CSEA violations in Gemini.' Google stated that '[t]hese tools are based on, or are similar to, the hash matching and machine learning classifiers that Google uses to detect CSEA across other Google services'.	Google added that it continues to improve the models that power Gemini to ensure that they respond safely to user prompts. It added 'Google also uses prompt classifiers to help determine whether or not a prompt is on a topic that could lead to an answer which violates our policies, and we leverage that to determine whether we should block the prompt altogether, or to dynamically re-prompt Gemini to generate a safer response. This includes prompts that seek child sexual abuse material or may result in violent Content.'
TVE	No[128]	Google stated that it 'deploys an internally developed machine learning classifier that has been trained to identify policy violations in user uploaded images, which includes TVE content.'	

C. Scanning outputs to detect AI generated TVE and CSEA

Google was asked about the use of automated tools to scan the outputs from Gemini to detect potential synthetic TVE and CSEA. Google provided the following information:

Table X

Harm type	Outputs scanned?	Details provided
TVE	Yes	• Response classifiers – Google uses classifiers to 'review the output of the models that power Gemini, and to block unsafe outputs before they are presented to the user, which includes outputs that would meet the definition of TVE.' • Output monitoring – Google deploys 'monitoring tools' which evaluate samples of Gemini outputs returned to a user and then use an algorithm to flag suspect outputs to be reviewed by a human to confirm whether they violate policy. The results of these evaluations are used to improve model responses.

[127] Google stated that 'In some cases, Gemini may be blocked from responding to a query that may include blocklist terms (and refuse to generate outputs).'
[128] Google stated that, 'In some cases, Gemini may be blocked from responding to a query that may include blocklist terms (and refuse to generate outputs).'

CSEA	Yes	• Response classifiers - as above. • Output monitoring – as above. • Stand-alone classifier for sexually-explicit material – In addition to the tools described for detecting TVE material, Google also reported that it uses a 'separate standalone classifier trained to identify sexually explicit responses, which includes CSEA.'

Google stated that responses that do not pass the response classifiers are blocked before they can be returned to the user.

D. User reporting of outputs containing AI generated TVE and CSEA

Google was asked if users could make 'in service' reports if their prompts in Gemini generated CSEA or TVE material. Google reported that end-users could make such 'in-service' reports about generated TVE and CSEA material without being required to locate a separate webform or email address. Google stated that every Gemini response to a prompt is accompanied by '"thumbs up"/"thumbs down" buttons' that allow end-users to give feedback about the content generated. Google said that end-users could leave feedback in a comment box and tag a 'thumbs down' with these three categories:

- 'Offensive/unsafe'
- 'Not factually correct'
- 'Other'

Google also said that end-users can select 'Report Legal Issue' to be directed to a webform that acts as 'Google's central content reporting tool.[129]' Users must then select 'Gemini' on this form to be taken to another form where they can submit a report. Google reported that '[a]ll problematic content-related requests are reviewed by specialist reviewers within Google's Trust & Safety team, and actioned appropriately'.

i. Action taken in response to a report

In response to a question about the action taken when Google received a report of AI generated synthetic TVE or CSEA, Google provided the following information:

[129] Legal Help, Report Content on Google, accessed 24 July 2024, URL: https://support.google.com/legal/troubleshooter/1114905?sjid=12316864348266639473-EU#ts=1115658. URL supplied by Google.

Table Y

Harm type	Action taken in response to 'Thumbs Down feedback'	Action taken in response to 'Report Legal Issue' webform
TVE	• Report is analysed by Google's Trust and Safety team. • If reported content is found to violate a policy (including TVE content policies) it will be tagged. • Action will be taken to mitigate the risk of Gemini behaving in a similar manner in future, e.g., to prevent Gemini from responding to similarly problematic prompts and/or block Gemini from producing similarly problematic outputs.	• 'LCPS agents' review the report and assess whether to accept or reject the removal request, or seek further information. • When LCPS takes down content, the report will be routed to Google's Trust and Safety team to prevent Gemini from producing similarly problematic responses in future.
CSEA	• Similar to process outlined for TVE, but synthetic content related to child safety (e.g., CSEA material) is routed to, reviewed and actioned by Google's team of child safety specialists. • Confirmed CSEA content is also reported to NCMEC.	• Similar to the process outlined for TVE, but LCPS will route the user report to the Child Safety team. • Confirmed CSEA content is also reported to NCMEC.

ii. Number of reports Google received about synthetically generated TVE and CSEA

Google was asked to report on the number of reports it received about synthetic TVE and synthetic CSEA generated by Gemini between 1 April 2023 – 29 February 2024. Google provided the following information:

Table Z

Harm type	Number of user reports
TVE	258 (reviewed under Gemini's Dangerous Content policies – which includes TVE content)
CSEA	86 (reviewed by Google's child safety team)

In response to a follow-up question from eSafety Google was unable to confirm the number of reports that resulted in confirmation that TVE and CSEA had been generated on Gemini.

E. Excluding harmful material from training data

i. Filtering 'high risk content' and ensuring training data is 'sufficiently high quality'

eSafety referred in the notice to the fact that, in its 2023 AI Principles Progress Update, Google had stated that 'training data was filtered for high-risk content and to ensure all training data is sufficiently high quality' and that 'Quality filters were applied to all datasets used to train the pre-trained Gemini Pro model. Safety filtering was applied to remove harmful content'.[130]

Google was asked to specify how it defined 'high-risk content', outline the criteria it used to determine that training data was 'sufficiently high quality', and to describe the 'quality filters' and tools it used to achieve this.

ii. Defining 'high risk' content

Google stated that it filtered training data to remove certain types of content. Google further stated that with the exception of certain types of content, such as CSEA, Google will not remove all forms of potentially objectionable or harmful content from an AI model's training data set, so that the model can recognise, identify and respond appropriately to new problematic content. This has been demonstrated in independent studies.[131]

iii. 'Determining 'sufficiently high quality' training data

In response to a question about what determined 'sufficiently high quality' training data, Google stated that in training AI models this depends on what the model is trying to achieve, but 'Within the industry (and Google) "high quality" is generally understood as data that is at least:

- Accurate, up-to-date, and relevant to what the AI model is seeking to achieve.
- Representative. The data must be sufficiently representative and diverse to address all possibilities that the AI model may encounter. Gaps, biases and stereotypes in training data can result in a model reflecting those in its outputs as it tries to predict a plausible response.
- Clean. This requires pre-processing of the data to remove errors, inconsistencies, and duplications that can introduce "noise" into the training process and degrade results'.

iv. Process for applying 'quality filters' to training data

Google stated that in order to achieve 'sufficiently high quality' training data sets for Gemini, it takes the following steps:

---

[130] Google, 'AI Principles Progress Update' 2023, accessed 29 January 2024, 'AI Principles Progress Updated 2023', accessed 29 January 2024, URL: https://ai.google/responsibility/principles/
[131] URL provided by Google. arXiv, 'A Pretrainers Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity', accessed 24 July 2024, URL: https://arxiv.org/abs/2305.13169

1. '[C]areful curation or selection of the data-sets to be used as part of the training corpus'
2. Apply 'a combination of heuristic rules and model-based classifiers to ensure that the data is of "sufficiently high quality"'
3. Apply safety filters to remove 'high risk content'
4. '[S]earch for and remove any evaluation data that may have been in its training corpus before using data for training'
5. Determine final data mixtures and weights 'through ablations on smaller models'
6. '[A]lter the mixture composition during training – increasing the weight of domain-relevant data towards the end of training.'

Google also noted that 'This process of data selection and refinement to identify the optimal data-set is an ongoing process that Google will seek to refine and improve over time.'

v. Using tools to exclude TVE and CSEA material from training data

Google was asked to report on the steps it took to ensure that TVE and CSEA material was excluded from the datasets used to train the machine learning models that underpin Gemini. Google provided the following information.

vi. Filtering known TVE and CSEA material

Table AA

Harm type	Tools used to exclude material from training data?	Details provided
TVE	Yes	• Filtering URLs to 'known violative content' and 'recent removals of violative content'. • In response to a follow-up question, Google provided the names of the internal proprietary tools it uses, as well as SafeSearch, which Google said aimed to filter 'some visual depictions of explicit sexual content and violence or gore'.
CSEA	Yes	• Hash-matching and classifiers are used to detect and remove CSAM, 'as well as a broader range of content including content that sexualises minors and pornography'.* • Filtering URLs known to link to CSAM. * In response to a follow-up question, Google clarified that it used CSAI Match to remove known CSEA videos, and an internally developed tool to remove known CSEA images. Google also clarified that it used classifiers to filter potential new and known CSEA from training datasets.

vii. Filtering new ('unknown') TVE and CSEA material

Table BB

Harm type	Tools used to exclude material from training data?	Details provided
TVE	Yes	• See Table AA
CSEA	Yes	• See Table AA

F. Off-platform monitoring to discover exploitable vulnerabilities

Google was asked to provide details of any forms of off-platform monitoring it used to alert itself when an end-user discovered and shared an exploitable vulnerability relating to Gemini.

Google responded by referring to the information it provided in response to questions about its use of off-platform monitoring to detect and anticipate TVE-related threats.

G. Red-teaming

In response to questions about red-teaming, Google stated that it understood red-teaming to mean

> the practice of identifying safety risks and vulnerabilities in the Gemini system by stepping into the role of an adversary and executing simulated attacks to test defences and operational response capabilities.

Google stated that it did undertake red-teaming of Gemini's outputs specific to both TVE and CSEA.

Google stated that it undertakes red-teaming of its AI foundational models to ensure they meet 'baseline safety performance.' Google also reported that it tests its generative AI products (including Gemini) before launch and periodically afterwards. Google added that it will red-team an already launched product when new features or functionality is added, or the underlying model is retrained or updated.

In response to a question about the solutions Google put in place to rectify any vulnerabilities identified during red-teaming, Google stated

> Depending on the identified issues, Google may make changes to address or correct the vulnerability identified before release.

In response to a follow up question from eSafety, Google clarified that when 'CSEA related violations' were identified through red-teaming of Gemini outputs during the report period

Google responded by adding 'relevant blocklists of terms'[132]. Google noted that none of these CSEA-related violations constituted 'reportable CSAM'. Regarding addressing TVE-related vulnerabilities during red-teaming, Google stated:

> During the Reporting Period, red-teaming of Gemini did not identify vulnerabilities related to TVE, indicating that no changes were needed.

H. Internal and external red-teaming

Google reported that it performed internal and external red-teaming of Gemini during the report period.

i. Internal red-teaming

Google stated that it used 3 specialist teams to red-team Gemini during the report period:

- Google's engineering team – Included work on trialling adversarial queries to attempt to trick Gemini into 'behaving badly.'
- Google's trust and safety teams – Included 'adversarial evaluations of stable sets of data' to compare responses to topics such as violent extremism.
- Google's child safety team – Involved adversary testing by child safety subject matter experts 'in a controlled environment, with appropriate and secure protections, to attempt to "break the model".'

ii. External red-teaming

Google stated that it facilitated external evaluations of Gemini Ultra (model) and Gemini Advanced (end-to-end product) in 2023. For Gemini Ultra, Google said that red-teaming candidates were selected based on their expertise and allowed to design their own testing methodology and prompt sets and wrote their reports independently of Google. For Gemini Advanced, Google said it used three types of external testing:

- Priority user program – 120 'power users, key influencers and thought-leaders' who focused on 'safety and persona, functionality, coding and instruction capabilities, and factuality'.
- Power users testing – 50 'power users' recruited through external vendors.
- Security testing – external testers with security backgrounds who conducted 'security and prompt-injection testing, jailbreaking, and user-interface security failures.'

[132] Google stated that 'In some cases, Gemini may be blocked from responding to a query that may include blocklist terms (and refuse to generate outputs).'

I. Penetration testing

In response to a question about the specific kinds of penetration testing Google conducted on Gemini's model, Google provided the following information:

Table CC

Type of penetration testing	Performed Yes/No	Details provided

Testing if the model is capable of producing images/symbols associated with designated terrorist organisations	Yes	• Google created 'evaluation datasets' consisting of thousands of 'adversarial' text and image prompts designed to stress test the model's capacity to generate different types of '"unsafe" content'.
		• It said these 'may include prompts designed to elicit TVE-type content'.
		• Google used these evaluation datasets to perform:
		  ○ Standalone classifier evaluations – which directly evaluate the classifier's ability to detect harmful prompts and responses.
		  ○ End-to-end product evaluation of Gemini – Evaluates the performance of its safety mechanisms in the context of Gemini's responses, including how often Gemini generates policy violations after protections are implemented.

Testing if the model is capable of producing content that is associated with CSEA	Yes	• Testing for CSEA is similar for TVE, but with added safeguards including being undertaken exclusively by specialist teams in a controlled and secure environment, compliant with all applicable laws.
Testing if the model would refuse certain instructions such as production of images/symbols associated with designated terrorist organisations and CSEA	Yes	• Google referred to its response to questions about testing of the model's capabilities regarding the generation of TVE and CSEA material.

J. Purple-teaming

In response to a question, Google stated that it did not perform **purple/violet-teaming**[133] of Gemini's outputs specific to TVE and CSEA during the report period.

[133] A collaborative approach to penetration testing where adversarial (red team) and defensive (blue team) teams work together to probe, refine, and strengthen defences against realistic simulated attacks.

Meta summary

Overview

Meta Platforms Inc was asked about three services it provides: Facebook, Messenger, and Instagram (including Threads).

1. Questions about Meta's definitions of 'terrorist material and activity' and violent extremist material and activity'

A. Terrorist material and activity

In response to a question about how Meta defines 'terrorist material and activity' or a different but equivalent term for the purposes of its terms of service and community guideline, Meta stated that on both Facebook and Instagram, such material and activity is covered by the 'Dangerous Organisations and Individuals' (DOI) section of the Facebook Community Standards. Meta specified that the Facebook Community Standards apply to Instagram in addition to the Instagram Community Guidelines.

Meta stated that, at a high level it aimed to remove glorification, support and representation of dangerous organisations and individuals. It defined dangerous organisation or individuals 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 organization;
- in order to achieve a political, religious or ideological aim.

Meta also noted that its definition is 'agnostic to the ideology or political goals of a group or individual' and that the test is 'whether they use violence to pursue those goals'. Meta reported that under the DOI policy, it designates and bans individuals and organisations tied to:

- terrorism;
- organised hate and large-scale criminal activity;

- multiple-victim violence and attempted multiple victim violence;
- serial murders;
- violent events;
- militarised social movements;
- violent non-state actors; and
- violence-inducing conspiracy networks such as QAnon.

Meta further stated that 'terrorist material and activity' is also covered by the:

- 'Violence and Incitement' section of the Facebook Community Standards, which 'prohibits content that incites or facilitates violence and constitutes a credible threat to public or personal safety'; and
- 'Coordinating Harm and Promoting Crime' section of the Facebook Community Standards, which 'prohibits users from facilitating, organizing, promoting, or admitting to certain criminal or harmful activities targeted at people, businesses, property, or animals'.

B. Violent extremist material and activity

In response to questions about how Facebook and Instagram define 'violent extremist material and activity' or a different but equivalent term for the purposes of its terms of service and community guidelines, Meta referred to the response it provided to eSafety's question about how it defines 'terrorist material and activity'.

2. Thresholds/criteria to determine action on TVE breaches

Meta was asked if it had criteria or thresholds in place to determine what action would be taken when TVE was identified on Facebook and Instagram. Meta provided the following information:

Table A

Actions taken on accounts or content when TVE was identified	Criteria/thresholds reported for Facebook and Instagram
Permanent account or user ban	Meta stated that it will disable a user for severe violations of its TVE policies, 'such as representing a designated organisation through profile name, photo, or description'.
Temporary suspension	Meta stated that for less serious violations, users will be temporarily restricted from some features as follows: • Two to six strikes: A user will be restricted from some features, such as posting in groups, for a limited amount of time. • Seven strikes: A user will get a 1-day restriction from creating content, which includes posting, commenting, creating a page and more. • Eight strikes: A user will get a 3-day restriction from creating content. • Nine strikes: A user will get a 7-day restriction from creating content. • Ten or more strikes: A user will get a 30-day restriction from creating content.
Account strikes	Meta stated that users will accrue strikes for violations of Meta's policies.
Blackholing of content	Meta stated that for most violations of Meta's DOI policies, 'the URL will be blackholed (blocked)'.
De-prioritisation in recommender system	Meta stated that if a user has posted content which does not violate Meta's policies, but which is covered by its recommendation guidelines[134], it will not be eligible for recommendation.

Meta noted that 'it is difficult to reduce the complexity of our enforcement policies into a single response' and that 'as a general rule our enforcement policies are designed to be proportionate, effective, and fair'.

[134] Facebook, 'What are recommendations on Facebook', URL: https://www.facebook.com/help/1257205004624246. URL supplied by Meta on 30 August 2024. Meta also supplied a URL that returns the Instagram log-in page, URL: https://help.instagram.com/313829416281232/?helpref=related_articles. URL supplied by Meta on 30 August 2024.

3. Questions about reporting of TVE

A. In-service reporting of TVE to Meta

In response to questions about whether users could report instances of TVE to Meta within its services (as opposed to navigating to a separate webform or email address), Meta responded:

Table B

Parts of service with an in-service reporting option
Facebook • Facebook Newsfeed • Facebook Groups • Facebook Channels • Facebook Stories **Messenger** • Messenger (when E2EE enabled) • Messenger (when E2EE not enabled) • Messenger Channels • Messenger Stories **Instagram** • Instagram Feed • Instagram Direct (when E2EE enabled) • Instagram Direct (when E2EE not enabled) • Instagram Groups • Instagram Reels **Threads** • Threads
Parts of service with no in-service reporting option
Messenger • Messenger Rooms

Meta identified the specific reporting categories set out in Table B below that users should pick to make a report of TVE (or close equivalent) for the relevant parts of the service.

Table C

Service	Part of service	Category used to report TVE in-service
Facebook	• Facebook Newsfeed • Facebook Groups • Facebook Stories	Terrorism
	• Facebook Channel	"Sharing Inappropriate Things" -> "Violent or Graphic content"
Messenger	• Messenger (when E2EE enabled) • Messenger (when E2EE not enabled) • Messenger Channels	"Sharing Inappropriate Things" -> "Violent or Graphic content"
	• Messenger Stories	Violence
Instagram	• Instagram Feed • Instagram Direct (when E2EE enabled) • Instagram Direct (when E2EE not enabled) • Instagram Groups • Instagram Reels	Violence or dangerous organisations
Threads	• Threads	Violence or dangerous organisations

In response to why there was no in-service reporting option for Messenger Rooms, Meta stated that

> In order to protect the privacy of our users and to comply with applicable law, including the U.S Wiretap Act, we do not record calls made via Messenger. As a result, if a user made a report about the content of a call made via Messenger, we would not be able to investigate that report as we do not have a record of that content.

Meta stated that users can report the relevant message thread of a Messenger Room using 'in service' reporting tools.

B. Reporting mechanisms for other entities to report TVE

In answer to a question about having separate reporting mechanisms for other entities to report TVE, Meta responded that Facebook and Instagram have reporting mechanisms (separate from users in general) for:

- law enforcement,
- Trusted Flaggers,
- regulatory and public authorities, and

- civil society groups.

Meta stated that reports made via these channels allow the reporting entity to provide additional context and/or evidence, which can assist with investigation and prioritisation of the report. The reporting entity is also provided a tracking number for their report and following review, the reporting entity is also informed of what action was taken.

4. Questions about proactive detection

A. Detecting known material using hash matching

i. Known TVE images

In response to questions about hash matching for known TVE images, Meta provided the following information:

Table D

Parts of service where hash matching tools are used for known TVE images	Names of tools used[135]
Facebook • Facebook newsfeed posts, including comments sections • Facebook profile pictures • Facebook Groups profile pictures • Facebook Groups (public) posts, including comment sections • Facebook Group (closed/private) posts, including comment sections • Facebook Channels • Facebook Stories **Instagram** • Instagram Feed • Instagram Direct (when E2EE not enabled) • Instagram profile pictures • Instagram Groups profile picture • Instagram Groups • Instagram Reels **Threads** • Threads • Threads profile picture	• SimSearchNet++ • PhotoDNA • PDQ

[135] Meta initially reported that Media Match Service was the name of the tool that it used to detect known TVE images. In response to a follow up question where eSafety noted that Media Match Service is not the name of a hash matching tool as defined in the Notice, Meta provided the names of the hash matching tools it used on the parts of its service.

Messenger • Messenger (when E2EE not enabled) • Messenger Group cover photos • Messenger Channels • Messenger Stories	• PhotoDNA • PDQ
Parts of service where hash matching tools are not used for known TVE images	
Instagram • Instagram Direct (when E2EE enabled) **Messenger** • Messenger (when E2EE enabled)	

ii. Known TVE videos

In response to questions about hash matching for known TVE videos, Meta provided the following information:

Table E

Parts of service where hash matching tools are used for known TVE videos	Names of tools used
Facebook • Facebook Newsfeed posts, including comment sections • Facebook Group (public) posts, including comments sections • Facebook Group (closed/private) posts, including comment sections • Facebook Channels • Facebook Stories	• Proprietary Meta video hashing tool • VideoMD5
Messenger • Messenger (when E2EE not enabled) • Messenger Channels • Messenger Stories	• Proprietary Meta video hashing tool
Instagram • Instagram Groups • Instagram Direct (when E2EE not enabled)	• Proprietary Meta video hashing tool • VideoMD5
Instagram • Instagram Feed • Instagram Reels	• Proprietary Meta video hashing tool • VideoMD5 • VideoPDQ
Threads • Threads	

Parts of service where hash matching tools are not used for known TVE videos
Messenger • Messenger (when E2EE enabled) **Instagram** • Instagram Direct (when E2EE enabled)

iii. Known TVE written material

In response to questions about hash matching for known TVE written material, such as manifestos or text promoting, inciting, or instructing in TVE, Meta provided the following information:

Table F

Parts of service where tools are used for known TVE written material	Names of tools used
Facebook • Facebook Newsfeed posts, including comment sections • Facebook Groups (public) posts, including comment sections • Facebook Group (closed/private) posts, including comment sections • Facebook Channels **Messenger** • Messenger (when E2EE not enabled) • Messenger Channels **Instagram** • Instagram Feed • Instagram Direct (when E2EE not enabled) • Instagram Groups **Threads** • Threads	Nilsimsa
Parts of service where hash matching tools are not used for known TVE written material	
Messenger • Messenger (when E2EE enabled) **Instagram** • Instagram Direct (when E2EE enabled)	

iv. Reason why tools are not used to detect known TVE on E2EE-enabled parts of service

In response to why hash matching tools are not used to detect known TVE images on Instagram Direct when E2EE is enabled, Meta stated, 'It is not technically possible to use hash-matching tools on [the end-to-end] encrypted parts of the service. However, we plan on rolling out the use of hash matching tools to detect known TVE images in group cover photos and in reported messages (which are not [end-to-end] encrypted) on Instagram Direct soon.'

eSafety notes that Meta has been working on the end-to-end encryption of Instagram Direct since at least 2019, when it was announced. eSafety considers that a key principle of Safety by Design, and the Expectations, is that safety should be built into a service or feature at the outset, rather than added later.

In response to why it did not have any measures in place to detect known TVE videos and written material in E2EE-enabled parts of Messenger and Instagram Direct, Meta repeated the technical obstacles outlined above.

With respect to these parts of the service, Meta reported that users can report TVE in Messenger and Instagram Direct, which is then used as a trigger for human review.

v. Sources of TVE hashes

Meta reported that it sourced its hashes of known TVE images and videos from the following databases:

- Meta's internal hash list generated from its experience reviewing content; and
- the Global Internet Forum to Counter Terrorism's (GIFCT) repository of hashes.

Meta stated that it updates its internal hash list depending on the frequency with which it identifies content eligible for banking. Meta stated that it updates its internal hash list automatically and in near real-time for the GIFCT repository.

For its hashes of known written TVE material, Meta said its databases were 'internally developed and manually curated' by its Dangerous Organizations and Individuals team and that the content is 'sourced from our own ongoing integrity work, as well as from investigations by paid third party vendors'.

B. Detecting New TVE material

i. New or 'unknown' images

In response to questions about the detection of new (or 'previously unknown') TVE images, Meta provided the following information:

Table G

Parts of service where tools are used for new TVE images	Names of tools used
Facebook • Facebook Newsfeed posts, including comment sections • Facebook profile pictures • Facebook group profile pictures • Facebook Group (public) posts, including comment sections • Facebook Group (closed/private) posts, including comment sections • Facebook Channels • Facebook Stories **Messenger** • Messenger Channels • Messenger Stories **Instagram** • Instagram Feed • Instagram profile pictures • Instagram Groups profile picture • Instagram Groups • Instagram Reels **Threads** • Threads • Threads profile picture	Unified Content Model
Parts of service where tools are <u>not used</u> for new TVE images	
Messenger • Messenger (when E2EE not enabled) • Messenger (when E2EE enabled) **Instagram** • Instagram Direct (when E2EE not enabled) • Instagram Direct (when E2EE enabled)	

ii. New or 'unknown' videos

In response to questions about the detection of new (or 'previously unknown') TVE videos, Meta provided the following information:

Table H

Parts of service where tools are used for new TVE videos	Names of tools used	Whether tools are video and/or audio classifiers, or others
Facebook • Facebook Newsfeed posts, including comment sections • Facebook Group (public) posts, including comment sections • Facebook Group (closed/private) posts, including comment sections • Facebook Channels • Facebook Stories **Messenger** • Messenger Rooms • Messenger Channels • Messenger Stories **Instagram** • Instagram Feed • Instagram Groups • Instagram Reels **Threads** • Threads	Unified Content Model	Text, image, video, and audio
Parts of service where hash matching tools are <u>not used</u> for known TVE videos		
Messenger • Messenger (when E2EE not enabled) • Messenger (when E2EE enabled) **Instagram** • Instagram Direct (when E2EE not enabled) • Instagram Direct (when E2EE enabled)		

iii. Text analysis to detect TVE

In response to questions about technology used to detect phrases, codes, and hashtags indicating likely TVE in text (for example manifestos or text promoting, inciting, instructing TVE) Meta provided the following information:

Table I

Parts of service where tools are used for phrases, codes, hashtags indicating likely TVE	Names of tools used
Facebook • Facebook Newsfeed posts, including comment sections • Facebook username • Facebook profile description • Facebook Group username (public and closed/ private) • Facebook Group profile description (public and closed/private) • Facebook Group (public) posts, including comment sections • Facebook Group (closed/private), including comment sections • Facebook Channels • Facebook Stories **Messenger** • Messenger Channels • Messenger Stories **Instagram** • Instagram Feed • Instagram username • Instagram user bio • Instagram Groups • Instagram Groups username • Instagram Groups profile description • Instagram Reels **Threads** • Threads • Threads bio	Unified Content Model
Parts of service where tools are <u>not used</u> for phrases, codes, hashtags indicating likely TVE	
Messenger • Messenger (when E2EE not enabled) • Messenger (when E2EE enabled) **Instagram** • Instagram Direct (when E2EE not enabled) • Instagram Direct (when E2EE enabled)	

iv. Reason tools are not used to detect new TVE on E2EE-enabled and other parts of services

In response to why it did not have any measures in place to detect new TVE images and videos, or to scan for indications of likely TVE in text, in E2EE-enabled parts of Messenger and Instagram Direct Meta stated that it was not technically possible to use classifiers, or to search for phrases, codes or hashtags on the end-to-end encrypted parts of the service.

In response to why it did not have any measures in place to detect new TVE images and videos, or to scan for indications of likely TVE in text, in parts of Messenger and Instagram Direct where

E2EE is not enabled, Meta stated it considers 'hash matching tools to be the most appropriate tool to detect TVE in private messaging threads'.

With respect to these parts of the service, Meta reported that users can report TVE in Messenger and Instagram Direct, which is then used as a trigger for review (automated or human).

v. Sources of phrases, codes, hashtags

Meta stated that its list of phrases, codes, and hashtags indicating likely TVE is 'manually curated by our Dangerous Organizations and Individuals team and sourced from our own ongoing integrity work, as well as from investigations by paid third party information vendors.'

C. Action taken on TVE

In response to questions about what action was taken when known TVE images, video, and written TVE material (known and new) was detected by its tools, Meta stated

> If a match is detected, the content is either automatically deleted or enqueued for human review. We may also take enforcement action at the account level.

Meta also stated that phrases, codes, or hashtags indicating likely TVE may be blocked.

For new TVE images and videos, Meta stated

> Depending on signals and confidence of the classifier, the content is either automatically deleted or enqueued for human review. We may also take enforcement action at the account level.

D. Livestreamed TVE

i. Detecting livestreamed TVE

Meta was asked to provide information about the measures it had in place to detect livestreaming on its service. The notice specified that livestreaming includes one-on-one video calls and video calls where one or more multiple people stream material to a group of any size.

Meta provided the following information:

Table J

Service	Measures in place to detect TVE in livestreams?	Interventions used	Names of tools used
Facebook Live **Instagram Live**	Yes	• Text classifiers • Video classifiers • Audio classifiers • Keywords • Behavioural signals	• Proprietary Meta video hashing tool • Proprietary Meta Classifier 1 • Proprietary Meta Classifier 2
Messenger Rooms	No	N/A	N/A

In response to why it did not have any measures in place to detect livestreamed TVE in Messenger Rooms, Meta stated:

> Meta differentiates between 'live streaming' products which are designed to enable a user to post a one-way broadcast of live events to large numbers or the general public and 'video calling' products which are designed to enable a user to have a private interpersonal end-to-end encrypted conversation with another user or a small group of users. While we implement a range of measures to detect live streamed TVE in our live streaming products, in order to protect the privacy of our users and to comply with applicable law, including the U.S. Wiretap Act, we do not proactively monitor private calls on video calling products like Messenger.

eSafety notes that it is concerning that Meta's Messenger does not detect livestreamed TVE given the use of Facebook Live in the Christchurch attack and Meta's public commitments (e.g. as part of the Christchurch Call) to take further steps to ensure the safety of its service. eSafety notes that Messenger Rooms enables up to 50 users to participate in livestream/live video at once.

ii. Reducing the likelihood of livestreamed TVE

In response to questions about the steps taken to reduce the likelihood that TVE could occur in livestreams, Meta stated that it used the following measures:

- Priority reviews of reports related to Facebook Live or Instagram Live – including prioritisation of livestream reports related to 'Violating Violent Events, above and beyond our prioritisation of Live video.

- Restrictions for those who have previously violated DOI policies for a set period of time starting from their first offence – for example, Meta stated that someone who 'shares a link to a statement from a terrorist group with no context will now be immediately blocked from using Live for a set period of time'.

Meta also stated that it banks content in its systems to prevent copies from being re-shared.

iii. In-service reporting of livestreams by users that are not logged into Facebook Live

In response to a question, Meta stated that there is no mechanism to enable users that are not logged-in to Facebook Live to make an in-service report about livestreamed TVE.

> eSafety notes that the inability for users not logged-in to Facebook Live to make an in-service report about livestreamed TVE may increase friction for users to report TVE, and prevent non-users from making reports at all. This is notable given the use of Facebook Live in the Christchurch attack.

In response to a question about the alternative steps Meta takes to ensure that its reporting mechanisms for livestreamed TVE are clear and readily identifiable (as expected by section 13 and 15 of the Determination), Meta stated

> The easiest way for a logged out user to report such material is to log back in and use our in-service reporting tools. However, we do offer a reporting tool[136] for logged out users to report violations'.

E. Languages covered by language analysis tools

In response to questions about the languages covered by Meta's language analysis tools, Meta stated that it uses the Unified Content Model to detect new TVE videos and phrases, codes, and hashtags indicating likely TVE in text. When asked about the languages covered by the Unified Content Model, Meta stated that the tool is language agnostic.

In response to follow-up questions from eSafety, Meta stated that the Unified Content Model involves two steps: text extraction and text analysis.

i. Text Extraction

Meta reported that 'text extraction can be done by audio transcription or optical character recognition (OCR)'. Meta reported that the list of languages covered by audio transcription are:

[136] Facebook, 'Report something on Facebook', URL: https://www.facebook.com/help/contact/485974059259751. URL submitted 30 August 2024. URL supplied by Meta on 30 August 2024.

Table K

Arabic	Bengali	Burmese	English	French	German
Hindi	Indonesian	Italian	Japanese	Kannada	Malay
Malayalam	Marathi	Portuguese	Russian	Sinhala	Spanish
Tamil	Thai	Turkish	Urdu	Vietnamese	

Meta reported that the list of languages covered by OCR are:

Table L

Amharic	Arabic	Bengali	Bulgarian	Burmese	Central Khmer
Chinese	Croatian	Dutch	English	French	German
Greek	Gujarati	Hebrew	Hindi	Hungarian	Indonesian
Italian	Japanese	Javanese	Kannada	Korean	Malay
Malayalam	Marathi	Persian	Polish	Portuguese	Punjabi
Romanian	Russian	Sinhala	Spanish	Tagalog	Tamil
Telugu	Thai	Turkish	Urdu	Vietnamese	

ii. Text Analysis

Meta reported that the Unified Content Model text analysis is done by a proprietary embedding algorithm that is pre-trained on the following languages:

Table M

Afrikaans	Albanian	Amharic	Arabic	Armenian	Assamese
Azerbaijani	Basque	Belarusian	Bengali	Bengali Romanised	Bosnian
Breton	Bulgarian	Burmese	Catalan	Chinese (Simplified)	Chinese (Traditional)
Croatian	Czech	Danish	Dutch	English	Esperanto
Estonian	Filipino	Finnish	French	Galacian	Georgian
German	Greek	Gujarati	Hausa	Hebrew	Hindi
Hindi Romanised	Hungarian	Icelandic	Indonesian	Irish	Italian
Japanese	Javanese	Kannada	Kazakh	Khmer	Korean
Kurdish (Kurmanji)	Kyrgyz	Lao	Latin	Latvian	Lithuanian
Macedonian	Malagassy	Malay	Malayalam	Marathi	Mongolian

Nepali	Norwegian	Oriya	Oromo	Pashto	Persian
Polish	Portuguese	Punjabi	Romanian	Russian	Sanskrit
Scottish Gaelic	Serbian	Sindhi	Sinhala	Slovak	Slovenian
Somali	Spanish	Sundanese	Swahili	Swedish	Tamil
Tamil Romanised	Telegu	Telegu Romanised	Thai	Turkish	Ukrainian
Urdu	Urdu Romanised	Uyghur	Uzbek	Vietnamese	Welsh
Western Frisian	Xhosa	Yiddish			

F. Blocking links to TVE material

i. Detection and sources of URLs

Meta was asked about its use of lists or databases to proactively detect and block URLs linking to TVE on other platforms. Specifically, Meta was asked about:

- Known URLs linking to websites/services operated by individuals/organisations dedicated to the creation, promotion, or dissemination of TVE or other TVE-related activities
- URLs linking to known TVE material on other services/websites (which may not be dedicated to TVE)
- Join-links to groups, channels, communities, or forums on other services that were known to be associated with TVE.

Table N

<table>
<tr><th colspan="3">Parts of service where Meta blocks URLs to:</th><th>URL sources</th></tr>
<tr><th>Websites/services dedicated to TVE</th><th>Known TVE material on other services/websites</th><th>Join-links to groups/channels on other services known to be associated with TVE</th><td></td></tr>
<tr><td colspan="3">Facebook
• Facebook Newsfeed posts, including comment sections
• Facebook profile description
• Facebook Group profile description (public and closed/private)
• Facebook Group (public) posts, including comment sections
• Facebook Group (closed/private), including comment sections
• Facebook Channels
Messenger
• Messenger (when E2EE not enabled)
• Messenger Channels
Instagram
• Instagram Feed
• Instagram Direct (when E2EE not enabled)
• Instagram bio
• Instagram Groups
• Instagram Groups profile description
Threads
• Threads
• Threads bio</td><td>Meta’s ‘own ongoing integrity work’ and investigations by paid third party vendors.</td></tr>
<tr><th colspan="4">Parts of service where Meta does not block URLs to:</th></tr>
<tr><th>Websites/services dedicated to TVE</th><th colspan="2">Known TVE material on other services/websites</th><th>Join-links to groups/channels on other services known to be associated with TVE</th></tr>
<tr><td colspan="4">Messenger
• Messenger (when E2EE enabled)
Instagram
• Instagram Direct (when E2EE enabled)</td></tr>
<tr><th colspan="4">Parts of service where Meta does not block URLs to:
Join-links to groups/channels on other services known to be associated with TVE</th></tr>
<tr><td colspan="4">Messenger
• Messenger Rooms - Meta reported that it is not possible to share URLs, including join-links, in Messenger Rooms</td></tr>
</table>

Meta reported that it uses an on-device functionality called 'Safe browsing' that enables it to detect URL snippets on Messenger and Instagram Direct when end-to-end encrypted messaging is enabled in order to 'warn users about potential issues with the links'. Meta stated that it sources these URL snippets from a 'database of harmful or suspicious URLs (including but not limited to URLs that may violate our DOI policies)'. Meta subsequently stated that the 'Safe browsing' feature is a user control, which users can turn on or off. Meta added that users can make reports about TVE in Messenger and Instagram Direct which will trigger human review.

ii. Action taken on accounts attempting to share blocked URLs/join-links

In response to questions about what action was taken when an account was detected attempting to share a blocked URL dedicated to TVE, a blocked URL linking to TVE on another website/service or a blocked join-link to groups or channels on other services known to be associated with TVE, Meta stated

> In general, users are blocked from sharing the URL and any existing posts or messages that include the URL will be removed. If the URL has been included in an existing "About me" section of a Facebook Page or "bio" section of an Instagram profile, the relevant Page or user will be prevented from being able to take certain actions until the URL is removed.

G. Off-platform monitoring

In response to a question about whether Meta used off-platform monitoring[137], either provided internally or by third-party services, to identify accounts, groups, channels, or communities dedicated to TVE on Facebook and Instagram, Meta stated it works with 'trusted external partners to help identify entities on Facebook and Instagram that may be involved in TVE' and uses this information as part of its process to designate dangerous organisations and individuals under its DOI policy. Meta stated that it designates dangerous organisations and individuals 'based on their behaviour both online and offline – most significantly, their ties to violence'.

Meta stated that it collects its information from, 'the GIFCT, Tech Against Terrorism, Global Network on Extremism and Technology and third party vendors. These third party vendors vary depending on operational needs. However, they generally include industry experts in online extremism, militant extremism, and foreign terrorist organizations'.

[137] Monitoring of activity on other services.

H. Percentage of reports sent for human review

In response to questions about the percentage of TVE reports sent for human review and the criteria and thresholds used to determine when reports were sent for review, Meta provided the following information:

Table O

	Percentage of user reports of TVE sent for human review	Criteria and thresholds used to determine when a user report is sent for human review	Percentage of TVE detected through automated tools sent for human review	Criteria and thresholds used to determine when a report of TVE detected through automated tools is sent for human review
Facebook	83.4%	• Severity – how severe the associated harm is with the reported content • Virality – how quickly the content in the user report is being shared • Likelihood of violation – where Meta has a signal and automation to help inform, how likely does the content in the user report violate Meta's policies	4.6%	Depends on the violation type and confidence level of the detection. Some violation types will be deleted immediately, others will be sent for review – including where an assessment of context is required. When Meta's classifiers detect violation signals, they generate a confidence score in likelihood of violation. If the confidence score is not high, the content may be sent for human review.
Messenger	39.7%		0.2%	
Instagram	87.8%		3.4%	
Threads	59.4%		3.2%	

Meta noted that these figures represent Australian user data for the period 1 October 2023 to 29 February 2024. Meta explained that this was because the data needed to distinguish between the relevant services was not consistently collected before this date.

I. Percentage of TVE detected proactively

Meta was asked what percentage of TVE was detected proactively, compared to TVE reported by users, trusted flaggers, or through other channels for the following services:

Table P

Service	Percentage of TVE detected proactively	Percentage of TVE reported by users, trusted flaggers or through other channels
Facebook Newsfeed	96.2%	3.8%
Facebook Groups (Public)	89.9%	10.1%
Facebook Groups (Closed/Private)	93.3%	6.7%
Messenger (E2EE and when E2EE not enabled)*	100%	0%
Instagram Feed	99.4%	0.6%
Instagram Direct (E2EE and when E2EE not enabled)*	100%	0%
Threads	93.2%	6.8%

Meta noted that these figures represent content created by Australian users that was removed due to TVE policy violations during the period 1 October 2023 to 29 February 2024. Meta explained that this was because the data needed to distinguish between the relevant services was not consistently collected before this date.

* Meta stated that it was unable to provide separate data for the E2EE and non-E2EE versions of Messenger and Instagram Direct because there was no way to reliably differentiate between end-to-end encrypted and non end-to-end encrypted message threads within Meta's enforcement datasets.

J. Appeals against TVE-related moderation

In response to a question about how many appeals were made by users for accounts banned or content removed for TVE, where the service was alerted by automated tools or user reports, and how many of those were successful, Meta provided the following information:

Table Q

How Meta was alerted to TVE	Number of appeals made for accounts banned for TVE breach	Number of appeals that were successful for accounts banned	Number of appeals made for material removed for TVE breach	Number of appeals that were successful for material removed
Facebook				
Automated tools	0.5K	0.3K	42K	3.4K
User reports	0.2K	0.1K	6.4K	0.6K
Instagram				
Automated tools	0.2K	0.1K	35K	2.9K
User reports	0.1K	0<100	0.7K	<100

Meta reported that these figures represent data from Australian accounts banned for violations of its TVE policies and content created by Australian users which was removed for violating its TVE policies.

5. Questions about resources, expertise, and human moderation

A. Trust and Safety

i. Trust and Safety and other staff

eSafety referred in the Notice to the fact that, in March 2023, Meta had announced reductions to its staffing numbers.[138] Meta was asked to provide the number of staff that were employed or contracted by Meta to carry out certain functions at the beginning and the end of the report period. Meta provided the following information:

Table R

Category of staff	31 March 2023*	31 December 2023*
Engineers employed by Meta focused on trust and safety	1,862	1,814
Content moderators employed by Meta**	0	0
Content moderators contracted by Meta	28,965	25,905
Trust and safety staff employed (other than engineers and content moderators)***	5,265	3,803

*** Meta reported that it could not provide staff data specific to the dates specified in the notice because it runs reports on its organisational numbers on a quarterly basis. Meta provided data as at 31 March 2023 and 31 December 2023 as an alternative.**

**** Meta reported that 'content moderators are generally employed by Meta's vendors'. Meta further reported that at 31 March 2023 there were 3,159 employees in its 'global operations team' and as at 31 December 2023 the figure was 1,967. Meta stated that its 'global operations team' focuses on 'work related to content moderation work (e.g., quality reviews, building protocols, managing contractors etc)'.**

*****Meta reported that this cohort included employees 'working in global operations and other non-engineering tech functions (i.e., product managers, researchers, designers, etc), legal, and policy'.**

Meta noted that the above figures do not include WhatsApp figures.

[138] Facebook, 'Update on Meta's year of efficiency', 14 March 2023, accessed 26 February 2024, URL: https://about.fb.com/news/2023/03/mark-zuckerberg-meta-year-of-efficiency

ii. Trust and Safety dedicated to minimising TVE

In response to a question about dedicated trust and safety team(s) responsible for minimising TVE on Facebook and Instagram, Meta reported that it had a 'core policy team specifically focussed on counter-terrorism and dangerous organisations'. Meta stated that this group includes, 'former academics who are experts on counterterrorism, former prosecutors and law enforcement agents, investigators and analysts, and engineers'. Meta stated that the team works to 'study trends in terrorism, organized hate, and other dangerous organizations and works with partners to better understand these organizations as they evolve'.

Meta provided the following information about the composition of its team:

Table S

Name of role/area of expertise	Number of staff	Number of contractors
Product and public policy experts	10	1

iii. Surge teams to respond to a TVE crisis

Meta was asked if it had a surge team(s) to respond to TVE crises, such as a livestreamed attack with content disseminated on Facebook, Instagram, or Messenger. Meta answered 'yes' and stated that it used a 'rapid response protocol' to respond to violent events, such as a livestreamed terrorist attack. Meta stated that members of its policy and operational teams are on call 24/7 to be able to deploy the protocol quickly in response to such events.

Meta stated that its Content Policy team assesses if an event is to be designated as a violating violent event under its Dangerous Organisations and Individuals Policy and its Operations team establishes whether there is any potential use of Live. Meta stated that if the event is designated, instructions are immediately issued to reviewers to remove any content containing 'glorification, support or representation (e.g., accounts belonging to the perpetrator) of the attack or the perpetrators, as well as perpetrator-generated content or bystander imagery showing the moment of attack on visible victims'.

Meta added that it also engages with members of the GIFCT to ensure that content such as livestreams can be hashed, shared and removed by members across multiple platforms.

B. Languages human moderators operate across

In response to a question about the languages that its human moderators operated across (both employees and contractors), Meta provided the following information:

Table T

Languages covered by both employees and contractors			
• **Albanian** • Amharic • Arabic • Armenian • Assamese • Azerbaijani • Bengali • Bosnian • Bulgarian • Burmese • Cantonese • Croatian • Czech • Danish • Dutch • English • Estonian • French • Georgian	• German • Greek • Gujarati • Hausa • Hebrew • Hindi • Hungarian • Indonesian • Italian • Japanese • Kannada • Kazakh • Khmer • Korean • Kurdish • Latvian • Lithuanian • Malay • Malayalam	• Mandarin • Marathi • Mongolian • Nepali • Oriya • Oromo • Pashto, Pushto • Persian • Polish • Portuguese • Punjabi • Romanian • Russian • Serbian	• Sinhala • Somali • Swahili • Swedish • Tamil • Telugu • Thai • Tigrinya • Turkish • Ukrainian • Vietnamese • Zulu

Table U

Languages covered exclusively by employees		Languages covered exclusively by contractors	
• Arabic (Gulf) • Arabic (Levant, Egypt, Iraq) • Arabic (Mahgreb) • Arabic (Sudan) • Bambara • Belarusian • Bemba • Bengali (India) • Czech and Slovak • Dari (Afghanistan) • Filipino • French (Sub-Saharan Africa)	• Fula • Igbo • Kirundi • Kituba • Lingala • Mauritian Creole • Norwegian • Sindhi (India) • Sindhi (Pakistan) • Spanish (Latin America) • Urdu (India) • Urdu (Pakistan) • Yoruba	• Afrikaans • Bhojpuri • Chhattisgarhi • Dari • Dhivehi • Finnish • Ganda	• Konkani • Lao • Luganda • Maltese • Marwari • Meitei • Mizo • Sindhi • Spanish (Castilian) • Tagalog • Tulu • Urdu • Uzbek

C. Median time to reach an outcome to user report of TVE

Meta was asked to provide the median time taken to reach an outcome[139] after receiving a user report about TVE for the following services:

Table V

Parts of the service	Reports from users globally	Reports from users in Australia
Facebook Newsfeed	6.5 hours	4.2 hours
Facebook Group (public)	6.7 hours	2.5 hours
Facebook Groups (closed/private)	0.8 hours	2 hours
Messenger (when E2EE enabled)*	0.1 hours	0.1 hours
Messenger (when E2EE not enabled)*	0.1 hours	0.1 hours
Instagram Feed	24.4 hours	15.5 hours
Instagram Direct (when E2EE enabled)*	4.3 hours	Meta reported that it did not have any reports from Australian users where content was determined to violate TVE policies.
Instagram Direct (when E2EE not enabled)*	5.8 hours	3 hours
Threads	56.3 hours	59.5 hours

*** Meta reported that it does not ordinarily track or report data that differentiates when E2EE is and is not enabled regarding response times to user reports on Messenger and Instagram Direct. Meta stated the data provided for these surfaces was 'sourced from non-core datasets and cannot be verified or validated'. It added that 'while Meta has sought to provide accurate data to the best of its ability, Meta has material concerns about the reliability of this data and considers that this data is not sufficiently robust to be used for further analysis.'**

Meta noted that these figures represent data from 1 October 2023 to 29 February 2024. Meta also reported that the figures were calculated by identifying all user reports on content that was confirmed to violate its TVE policies and 'calculating the 50th percentile of the times taken from the creation of a job to the time an enforcement action was taken'. Meta noted that the creation of a job is when 'a user report cannot be closed automatically (e.g. due to duplication).'

eSafety notes the significantly longer time to respond to TVE reports on Threads than on other Meta services/parts of services. It is unclear why Threads reports are responded to more slowly.

[139] Defined in the Notice as a calculation from 'the time that a user report is made, to a content moderation outcome or decision, such as removing the content, banning the account, or deciding that no action should be taken.'

D. Volunteer moderation

Table W

Question	Response
Did Meta have a standards policy, or similar, outlining the responsibilities and expectations of volunteer moderators?	Yes Meta stated that, 'like all Facebook users, Facebook group admins and moderators are subject to the Facebook Community Standards[140]. We provide guidance on understanding the Community Standards[141], on creating and enforcing group rules[142], and on managing difficult group members[143]. The Help Center[144] also contains general guidance on how to manage people and content in groups. Meta also stated that, 'We generally remove groups that repeatedly violate the Facebook Community Standards. This includes if an admin of a group creates content, such as posts, titles, or group rules that violate our Community Standards or if a group admin or moderator approves violating content from a group member'.
What training and/or guidance was provided to Meta volunteer moderators regarding proactive minimisation of TVE and removal of accounts that share TVE.	Meta reported that it provides guidance to group admins and moderators on understanding Community Standards, creating and enforcing group rules, and managing difficult group members. Meta also stated > While group admins and moderators have an important role to play in keeping their communities safe and engaged, we do not expect them to take the lead in handling TVE content, as doing so could put their safety and wellbeing at risk. We invest heavily in developing clear policies with subject matter experts and deploying specialist tools to detect and take action against violations of TVE policies.
Were users able to make in service reports about volunteer moderators in instances where they were failing to meet any required responsibilities and expectations?	Meta responded 'Yes' Meta's response indicated that a user can report the group in service. It did not indicate that a specific report about a volunteer moderator can be made in service

[140] Meta, 'Facebook Community Standards', URL: https://transparency.meta.com/en-gb/policies/community-standards/, URL supplied by Meta on 24 June 2024.
[141] Facebook, 'Understanding Community Standards', URL: https://www.facebook.com/community/using-key-groups-tools/understanding-community-standards/. URL supplied by Meta on 30 August 2024
[142] Facebook, 'Establishing Membership and Rules', URL: https://www.facebook.com/community/establishing-membership-and-rules/, URL supplied by Meta on 30 August 2024.
[143] Facebook, 'Managing difficult members', URL: https://www.facebook.com/community/establishing-membership-and-rules/. URL supplied by Meta on 30 August 2024.
[144] Facebook, 'Managing people and content', URL: https://www.facebook.com/help/1686671141596230?ref=hc_about&helpref=about_content URLs supplied by Meta on 30 August 2024.

If volunteer moderators removed an account from a Facebook group for TVE-related breaches, were trust and safety staff informed?	No Meta stated > While group admins and moderators have an important role in setting the expectations and norms for their groups, we do not expect them to have the expertise to handle TVE content. For this reason, we invest heavily in detecting and taking action on TVE material on our services, including in groups.
If Meta's Trust and Safety staff banned a user for a TVE-related violation in a Facebook group, were the volunteer moderators of that group notified?	No In response to a question about the alternative steps Meta took to ensure that volunteer moderators were alert to the potential increased risk of TVE in a group, Meta stated > An admin can refer to the Community Quality[145] tool to obtain an overview of the content that has been removed or flagged to them for violating certain Community Standards, including those relating to TVE. This tool gives admins more clarity about how and when we enforce our policies in their groups and gives them greater visibility into what is happening in their communities.

6. Questions about steps to prevent recidivism

A. Measures and indicators

Meta provided a response to questions about the measures it took to prevent recidivism for TVE-related breaches on Facebook, Messenger, Instagram, and Threads. eSafety has chosen not to publish all the information Meta provided to prevent the information being misused.

Meta stated

> We use a combination of human and automated review to enforce against recidivist profiles..., including a large portion of enforcement that occurs during account registration (to ensure we enforce upon accounts as soon as possible once we have high confidence in said connection).

Meta listed multiple indicators[146] to detect users who have previously been banned for TVE-related breaches. eSafety has chosen not to publish these indicators to prevent the information being misused.

[145] Facebook, 'Understanding Community Quality', URL: https://www.facebook.com/community/using-key-groups-tools/understanding-group-quality/ . URL supplied by Meta on 30 August 2024.

[146] eSafety uses the following terms to give an impression of the extent of the indicators used by services in the table below, rather than publishing the specific indicators which could be misused:
- Minimal: A small number
- Several: A moderate number
- Multiple: A significant number.

Meta stated that it does not use all indicators by default in instances where an account was banned to prevent recidivism by that user. Meta stated that the set of indicators used can vary based on the account and the method of prevention in question and that there is no fixed criteria that governs the use of each indicator. Meta stated that the set of indicators used changes over time, and data is regularly reviewed to improve performance.

Meta also reported that it reserves using certain specific anti-recidivism measures to 'only the most severe use cases' including 'users who have been disabled for certain severe violations of our DOI policies'.

B. Preventing group recreation after ban

In response to a question about the measures Meta took to prevent banned TVE groups from being recreated on Facebook and Instagram, Meta stated that it used the following measures:

- Strategic Network Disruptions – targeted at a banned group's presence across Meta's services, which can be used to 'disrupt an entire network at once' and 'send a clear message to the group that we are aware of their presence and they are not welcome on our platforms'.
- Signals – involves identifying 'signals that indicate a banned organization has a presence, and then proactively investigate associated accounts, Pages and Groups before removing them all at once'. Once Meta has removed the groups presence it works to 'identify attempts by the organization to come back on our platform.'
- Sweeps – 'We conduct ongoing enforcement sweeps against known bad actors to ensure they do not continue to abuse our platforms.'
- DOI banks – 'We have banked some DOI names so that any pages/groups created with the same name are disabled.'

C. Applying TVE-related bans to associated accounts

Meta was asked, when it took action against a user for a TVE-related breach, whether it applied bans to associated accounts. eSafety defined 'associated accounts' as 'other users who are associated with the banned user'. Meta stated

> We designate dangerous organizations and individuals based on their behavior both online and offline – most significantly, their ties to violence. As part of the designation process, we identify signals that indicate a banned organization has a presence on our platforms, and then we use technology to "fan out" and proactively investigate associated accounts, Pages, and Groups, before removing this "cluster" all at once.

D. Sharing of banned account details

i. Sharing banned account details between Facebook and Instagram

In response to questions asking whether Facebook and Instagram share details about accounts banned for TVE on their respective services, Meta stated that both services mutually share such information with each other in certain specific circumstances. In response to a request for clarification from eSafety, Meta subsequently stated that both services share information about accounts banned on one service to identify accounts belonging to the same end-user on the other, but only take action to ban other identified accounts in certain specific circumstances.

eSafety has chosen not to publish additional information about these circumstances to avoid this information being misused. Meta stated that propagating a ban on one service to the other was limited to cases where a reliable match can be established between the accounts on each platform because 'The decision to permanently disable an account is not one that is made lightly and we therefore need to be confident that an account is associated with a particular user before disabling it.'

ii. Sharing of banned account details with other entities

Meta was asked if Facebook and Instagram shared details of accounts banned for TVE with the following entities:

Table X

Entity	Shared details of accounts banned for TVE?	Details provided by Meta
WhatsApp	Yes	Meta stated it will share certain Facebook and Instagram information with WhatsApp 'for severe violations of our DOI and other relevant policies'. Meta also stated that 'this may not occur in relation to users located in certain jurisdictions due to local privacy and other compliance obligations'.
Other service providers (Non-Meta)	No	Meta stated 'we may share limited information related to threats to mitigate risk of cross-platform abuse'.
Law enforcement	Yes	Meta stated '[w]e may share information related to credible threat(s) of harm or in response to a valid request from law enforcement'.
Regulatory or other public authorities	No	N/A
Global Internet Forum to Counter Terrorism	No	N/A
Civil society groups	No	N/A

7. Questions about recommender systems

A. Preventing amplification of TVE

i. Recommender algorithm – interventions

In answer to a question about whether Meta had interventions in place to prevent the amplification of TVE via its recommender algorithms on Facebook and Instagram, Meta referred to the information it provided regarding the measures it takes to remove TVE from its services.

ii. Recommender algorithm – testing

In answer to a question about any testing Meta performs to ensure that its recommender systems do not amplify TVE, Meta reported that during the report period it had not performed any such testing on either Facebook or Instagram.

In response to why it did not have testing measures in place to mitigate instances of amplification of TVE on Facebook and Instagram, Meta stated

> As TVE is prohibited by the Facebook Community Standards and the Instagram Community Guidelines, our measures are focussed on removing that content from our services (rather than preventing its amplification).

iii. Recommender algorithm – positive interventions

Meta was asked if Facebook or Instagram had systems in place to stage positive interventions, for example by promoting deradicalising content for at-risk users when a user sought out TVE material on the service. Meta reported

> If a user in Australia searches on Facebook or Instagram using words associated with organized hate or violent extremism, the top search result will be a link to resources and support for how to leave violence and extremism behind. We partner with Step Together in Australia to provide these resources and support.

8. Questions about generative AI safety

A. Implementing Meta AI in Australia

Meta was asked if it had taken steps with the goal of implementing Meta AI in Australia during the report period, which had not been launched at the time the Notice was given. Meta reported that it had taken steps and stated

> Prior to its launch in Australia, Meta AI was reviewed by the Australian legal, policy and comms teams to identify any local risks or concerns associated with the launch. Meta AI was

also subject to red teaming efforts by the local team to test for unique local risks. This is in addition to the extensive risk assessments that were conducted at a global level.

B. Safety risk assessments regarding TVE and CSEA

Meta was asked if it had undertaken internal safety risk assessments during the report period regarding the risk of Meta AI generating TVE and CSEA prior to implementing Meta AI in Australia. Meta reported that,

> An internal risk assessment was conducted to evaluate the potential risks associated with Meta AI and put in place mitigations to reduce those risks. The assessment considered several categories of content risks, including child sexual exploitation and terrorism risks. However, our risk assessment process is ongoing and we will continue to evaluate and seek to mitigate the potential risks associated with Meta AI.
>
> In addition, Meta AI was subject to review by external and internal experts through red teaming exercises to find unexpected ways that Meta AI might be used (including TVE violations). We then addressed issues identified as part of risk mitigation or remediation prior to launch.

9. Questions about end-to-end encryption

A. Safety risk assessments regarding TVE

Meta was asked if it had undertaken internal safety risk assessments during the report period regarding its ability to detect and address TVE specifically before implementing E2EE on Messenger and Instagram Direct[147]. Meta reported that it did not.

Meta stated

> While Meta did not undertake a safety risk assessment specifically in relation to TVE *during the Report Period*, such a risk has been actively considered by Meta as part of its ongoing risk assessment process.
>
> As part of this process, Meta created dedicated safety teams across the company to understand how end-to-end encryption could impact on existing safety mitigations and to identify [end-to-end] encryption-resilient approaches where needed. Meta also had regular engagements with 400+ NGOs and industry experts, including those in the terrorism space,

[147] In December 2023 (during the report period) Meta publicly announced that it was implementing end-to-end encryption (E2EE) by default on one-to-one Messenger chats and calls. Meta also announced that it was planning to 'expand this work as well as conduct additional testing around E2EE on Instagram over the next year'.

> to identify key risks that may be affected by [end-to-end] encryption and how to mitigate them.

Meta also referred to two public reports it had commissioned concerning impacts, risks, and mitigations relating to E2EE:

- Tech Against Terrorism's '*Terrorist Use of E2EE: State of Play, Misconceptions, and Mitigation Strategies*'[148], published September 2021; and
- Meta's 'Meta Response: End-to-End Encryption Human Rights Impact Assessment'[149], published April 2022.

Meta stated that it had 'deployed 30+ [end-to-end] encryption resilient safety features since 2019 and is working on implementing more', and that it 'continues to monitor the impact of end-to-end encryption on safety risks, including TVE'.

B. Interoperable E2EE messaging

Meta was asked if had undertaken work during the report period on interoperable E2EE messaging between Messenger, Instagram Direct, and WhatsApp.

Meta stated that it had not.

However, Meta referred to its March 2024 blog post, 'Making messaging interoperability with third parties safe for users in Europe'[150] for further information about its plans for interoperable messaging between its services and third-party services.

10. Additional information provided by Meta

Providers were given the opportunity to provide any other relevant, specific information in relation to additional or alternative steps they were taking to comply with each of the Expectations as set out in their respective notices. Meta stated

> To help other platforms that may not have the resources and the technology, we have developed and made available[151] a free open source software tool called Hasher-Matcher-

[148] Tech Against Terrorism, 'Terrorist Use of E2EE: State of Play, Misconceptions, and Mitigation Strategies', September 2021, accessed 4 July 2024, URL: https://www.techagainstterrorism.org/hubfs/TAT-Terrorist-use-of-E2EE-and-mitigation-strategies-report-.pdf. URL supplied by Meta.
[149] Meta, 'Meta Response: End-to-End Encryption Human Rights Impact Assessment', April 2022, accessed 4 July 2024, URL: https://about.fb.com/wp-content/uploads/2022/04/E2EE-HRIA-Meta-Response.pdf. URL supplied by Meta.
[150] Meta, 'Making messaging interoperability with third parties safe for users in Europe', 6 March 2024, accessed 4 July 2024, URL: https://engineering.fb.com/2024/03/06/security/whatsapp-messenger-messaging-interoperability-eu/. URL supplied by Meta.
[151] Meta, 'Meta Launches New Content Moderation Tool as It Takes Chair of Counter-Terrorism NGO', 13 December 2022, accessed 4 July 2024, URL: https://about.fb.com/news/2022/12/meta-launches-new-content-moderation-tool/. URL supplied by Meta.

Actioner (HMA) that identifies copies of images or videos and takes action against them en masse. HMA builds on Meta's previous open source image and video matching software, and it can be used for any type of violating content.

Meta also stated that it publishes the efficacy of its efforts 'to reduce the prevalence of terrorist content on Facebook and Instagram' in its quarterly Community Standards Enforcement Reports[152].

[152] Meta, 'Dangerous Organisations: Terrorism and Organized Hate', accessed 4 July 2024, URL: https://transparency.meta.com/reports/community-standards-enforcement/dangerous-organizations/facebook/. URL supplied by Meta.

WhatsApp summary

Overview

WhatsApp LLC was asked about its WhatsApp service.

1. Questions about WhatsApp's definitions of 'terrorist material and activity' and 'violent extremist material and activity'

A. Terrorist material and activity

In response to a question about how WhatsApp defines 'terrorist material and activity' or a different but equivalent term for the purposes of its terms of service and community guidelines, WhatsApp referred to its WhatsApp Messaging Guidelines[153] and WhatsApp Channels Guidelines.[154] WhatsApp stated that these guidelines 'prohibits the use of its service for sharing or engaging in illegal activity', and provided different examples of such TVE-related material and activity that is prohibited on private messaging and channels:

i. Private Messaging

WhatsApp reported that terrorist material and activity includes:

- 'Content that supports designated terrorist organisations or individuals; and
- Content that organises or coordinates violent crimes or violence against others, such as content that constitutes a credible threat to public or personal safety.'

ii. Channels

WhatsApp reported that terrorist material and activity includes:

- 'Content that supports violent extremist or criminal organisations or individuals; and
- Content that could cause serious harm to people, such as content that constitutes a credible threat to public or personal safety, incitement of violence, organisation or coordination of violent or criminal activities.'

[153] WhatsApp, 'WhatsApp Messaging Guidelines', provided by WhatsApp LLC 13 May 2024, URL: https://www.whatsapp.com/legal/messaging-guidelines
[154] WhatsApp, 'WhatsApp Channels Guidelines', provided by WhatsApp LLC 13 May 2024, URL: https://www.whatsapp.com/legal/channels-guidelines/

B. Violent extremist material and activity

In response to a question about how WhatsApp defines 'violent extremist material and activity' or a different but equivalent term for the purposes of its terms of service and community guidelines, WhatsApp referred to the response it provided to eSafety's question about how it defines 'terrorist material and activity'.

2. Access to Meta's 'Dangerous Organisations and Individuals' list

eSafety highlighted in the Notice that WhatsApp's parent company, Meta has publicly stated that it maintains an internal list that designates organisations and individuals 'that proclaim a violent mission or are engaged in violence' and prohibits their presence 'on Meta'.[155]

In response to a question from eSafety, WhatsApp stated that it does not prohibit all organisations on this list for the **private messaging part of its service.** WhatsApp reported that organisations on specific terrorist lists such as the US Foreign Terrorist Organisations list, the US Specially Designated Global Terrorist List, and the US Specially Designated Narcotics Trafficking Kingpins list are prohibited from using WhatsApp's private messaging features.

WhatsApp reported that it prohibits all organisations on Meta's Dangerous Organisations and Individuals list from using **WhatsApp Channels.**

> eSafety notes that it is unclear why WhatsApp does not consider prohibiting the same organisations as Meta on its private messaging but does consider that these organisations should be prohibited on Channels. eSafety considers that this discrepancy may mean that TVE organisations are able to operate on parts of WhatsApp without action taken against them by the service.

3. Thresholds/criteria to determine action on TVE breaches

WhatsApp was asked if it had criteria or thresholds in place to determine what action would be taken when TVE was identified on WhatsApp. WhatsApp provided the following information:

[155] Facebook, 'Dangerous organisations and individuals', accessed 26 February 2024, URL: https://transparency.fb.com/en-gb/policies/community-standards/dangerous-individuals-organizations/

Table A

Actions taken on accounts or content when TVE was identified	Criteria/thresholds reported for WhatsApp
Permanent account or user ban	WhatsApp stated that it will permanently ban an account when the user is identified engaging in illegal activity such as the examples provided in the above definition section.
Account strikes	WhatsApp stated that in certain contexts it will apply an account strike as a form of graduated enforcement. It added that accumulation of a certain number of strikes will result in a permanent account ban.
Community/Group suspension	WhatsApp stated that it will suspend a group/Community that it determines represents a designated organisation. WhatsApp added that '[a] suspended Community or group can no longer operate on WhatsApp.'
Channel enforcement	WhatsApp stated that 'violating content in a Channel may result in enforcement against that Channel, including admin account bans, restrictions on Channel discovery, and Channel suspension.' WhatsApp reported that it will suspend Channels for severe TVE violations (e.g. a Channel that demonstrates representation of a designated organisation) and may ban the channel owner and/or admins. For less severe violations WhatsApp reported that it 'takes steps to ensure that the violating content is not further disseminated', and that it will take the following steps until the violating content is removed: • Channel is removed from discovery surfaces • No new followers are able to find the Channel.

4. Questions about reporting of TVE

A. In-service reporting of TVE on different parts of the WhatsApp service

In response to questions about whether users could report instances of TVE to WhatsApp within the service (as opposed to navigating to a separate webform or email address), WhatsApp responded:

Table B

Parts of the service	In-service reporting option?	Reporting category
Direct messages (including groups)	Yes	'Report'
Communities	Yes	
Channels	Yes	
Status	Yes	

B. Reporting mechanisms for other entities to report TVE

In answer to a question about having separate reporting mechanisms for other entities to report TVE, WhatsApp responded that it does have reporting mechanisms (separate from users in general) for:

- Law enforcement
- Trusted Flaggers
- Regulatory and public authorities, and
- Civil society groups

WhatsApp stated that, 'Reports made via these channels open up a direct line of communication between the reporting entity and Meta's operational teams and allow the reporting entity to provide additional context and/or evidence, which can assist with WhatsApp's investigation and prioritisation of the report.'

WhatsApp added that it is important for users to use WhatsApp's in-service reporting tools because

> if the potentially violating account or content is not also reported via WhatsApp's in-app reporting tools, any relevant behavioral signals and content may not be incorporated into WhatsApp's machine learning systems.

5. Questions about proactive detection

A. Detecting known material using hash-matching

i. Known TVE images

In response to questions about hash matching for known TVE images, WhatsApp provided the following information:

Table C

Parts of service	Used image hash matching tools?	Names of tools used
Channels messages	No – but since implemented.	N/A
User profile picture	Yes	Media Match Service*
Groups profile picture	Yes	Media Match Service*
Communities profile picture	Yes	Media Match Service*
Channels profile picture	No	N/A
Status	No	N/A
Content in user reports	Yes	Media Match Service*

***In response to a follow-up question about Media Match Service (MMS), WhatsApp stated that the Media Match Tool is a system that involves the following steps:**

- Content uploading/retroaction
- Hash extraction
- Search and match
- Action

WhatsApp stated that 'embedding algorithms', such as PhotoDNA, are subcomponents of the MMS tool that focus on hash extraction.

eSafety notes that Meta provided more details about the hash matching tools used within the Media Match Service in response to its Notice. See section Questions about proactive detection.

In response to why hash matching tools are not used on Channels, and whether alternative steps were taken to detect known TVE images, WhatsApp stated Channels is a 'relatively new product' and that 'WhatsApp is currently working on the rollout of hash matching tools for TVE on Channels and expects them to be in use soon.' WhatsApp also noted that all content on Channels is not E2EE and that classifiers are used.

In response to why hash matching tools are not used on an end-user's Status and whether alternative steps were taken, WhatsApp stated that 'A user's status on WhatsApp is [end-to-end] encrypted' and that it is not possible to use hash-matching on E2EE parts of the service. WhatsApp also noted that if a user's status is reported, WhatsApp will use hash-matching tools on the reported content.

ii. Known TVE video

In response to questions about hash matching for known TVE video, WhatsApp provided the following information:

Table D

Parts of service	Used video hash matching tools?	Names of tools used
Channels messages	No – but since implemented.	N/A
Status	No	N/A
Content in user reports	Yes	Media Match Service*

***In response to a follow-up question about Media Match Service (MMS) WhatsApp stated that MMS is a system (as outlined above in the 'known TVE images' section) and that there are subcomponents of the MMS tool that focus on hash extraction.**

In response to why hash matching tools are not used to detect known TVE videos on Channels messages and status, WhatsApp referred to its reasons for not using such tools to detect known TVE images.

eSafety notes that WhatsApp deployed a new feature, WhatsApp Channels, without implementing hash-matching tools to detect known TVE images and videos and reported that only during the report period did it start working on its implementation. WhatsApp subsequently advised eSafety that hash-matching tools for TVE on Channels have been deployed since May 2024 (some 10 months after WhatsApp Channels was introduced[156]). eSafety considers that a key principle of Safety by Design, and the Expectations, is that safety should be built into a service or new feature at the outset, rather than added later.

iii. Known TVE written material

In response to questions about hash-matching for known TVE written material on WhatsApp, such as manifestos or text promoting, inciting, or instructing in TVE, WhatsApp provided the following information:

[156] WhatsApp, 'Introducing WhatsApp Channels. A private way to follow what matters', 8 June 2023, accessed 18 September 2024, URL: https://blog.whatsapp.com/introducing-whatsapp-channels-a-private-way-to-follow-what-matters.

Table E

Parts of service	Used image hash matching tools for written material?
Channels messages	No
Content in user reports	No

In response to why hash-matching tools are not used to detect known TVE written material on channels messages and content in user reports, WhatsApp stated that it believes that 'our text-based classifiers are the appropriate tool to detect TVE written material'.

iv. Sources of TVE hashes

WhatsApp reported that it sourced its hashes of known TVE images and videos from the following databases:

- WhatsApp's own internal hash lists (images only)
- Meta's hash lists (images and video)

WhatsApp stated that it ingested all hashes from these databases, and that updates to the databases depends on the frequency it (or Meta, as applicable) identifies eligible material.

B. Detecting new TVE material

i. New or 'unknown' TVE images

In response to questions about the detection of new (or 'previously unknown') TVE images, WhatsApp provided the following information:

Table F

Parts of service	Used tools for images?	Names of tools used
User profile picture	No	N/A
Groups profile picture	Yes	CT Image Classifier* Whole Post Integrity Embeddings Service
Communities profile picture	No	N/A
Channels profile picture	Yes	CT Image Classifier* Whole Post Integrity Embeddings Service
Channels messages	Yes	CT Image Classifier* Whole Post Integrity Embeddings Service
Status	No	N/A
Content in user reports	No	N/A

***WhatsApp stated that it uses 'embedding algorithms' that are subcomponents of the CT image classifier tool.**

In response to why it does not use any automated tools to detect new TVE images on the identified parts of its service, and whether alternative steps were taken to detect new TVE images, WhatsApp stated the following:

- User profile picture: 'In WhatsApp's experience, user profile pictures do not represent a useful signal of likely violating TVE presence or activity on WhatsApp's platform. Many users source them from the internet, change them frequently, and do not use them to represent actual identity. Enqueuing accounts for human review based on these signals diverts resources from higher priority reviews'. WhatsApp also noted that it prioritised other surfaces that it considered were more reliable indicators of violating TVE presence or activity. Communities profile picture: 'Communities remains a relatively new feature, which is still gaining adoption. The announcement groups only offer one-way communication and not the full range of features of a WhatsApp group. WhatsApp has focused its attention and resources on group activity, including groups within communities.' WhatsApp also noted if a reviewer determines that a group within a Community is violating for TVE, they will 'look at the overarching community information, including the profile picture' to determine if the whole community is violative and should be suspended and the admin accounts banned.
- Status: 'A user's status is [end-to-end] encrypted. WhatsApp is unable to use technology to detect new TVE images on [end-to-end] encrypted parts of the service'.
- Content in user reports: 'In WhatsApp's experience, user reported content has been a relatively weak signal as compared to group metadata.' WhatsApp also stated that it uses classifiers to enqueue user reports for human review and that it is currently 'investing in building additional automated tools to review user reported content, including images'.

ii. New or 'unknown' TVE videos

In response to questions about the detection of new (or 'previously unknown') TVE videos, WhatsApp provided the following information:

Table G

Parts of service	Used tools for videos?	Names of tools used	Whether tools are video and/or audio classifiers, or other
Status	No	N/A	N/A
Channels messages	Yes	Whole Post Integrity Embeddings Service	Video and audio
Content in user reports	Yes	CT Text Classifier	Audio

In response to why it does not use any automated tools to detect new TVE videos on user statuses, and any alternative steps taken to detect new TVE video, WhatsApp stated that statuses are E2EE and it cannot use technology to detect new TVE videos on E2EE parts of the service. WhatsApp noted that if a user's status is reported it extracts text from the audio on videos reported and uses the classifiers listed above to determine if that text violates its TVE policies.

eSafety notes that it is not clear why WhatsApp uses tools to detect new TVE videos in user reports, but does not do so for new TVE images, particularly when it can combine its tools with human review.

iii. Text analysis to detect TVE

In response to questions about technology used to detect phrases, codes, hashtags, indicating likely TVE in text (for example manifestos or text promoting, inciting, instructing TVE), WhatsApp provided the following information:

Table H

Parts of service	Used text analysis tools?	Names of tools used
User profile depiction	No	N/A
Groups profile description	Yes	CT Text Classifier*
Communities profile description	Yes	CT Text Classifier*
Channels profile description	Yes	CT Text Classifier*, Whole Post Integrity Embeddings Service
Channels messages	Yes	CT Text Classifier*, Whole Post Integrity Embeddings Service
Status	No	N/A
Content in user reports	Yes	CT Text Classifier*

*** WhatsApp stated that it uses 'embedding algorithms' that are subcomponents of the CT text classifier tool.**

In response to why it does not use technology to scan user profile descriptions for indications of likely TVE, WhatsApp stated '[i]n our experience, the risk of phrases, codes, hashtags indicating likely TVE in text in user profile descriptions is low.'

In response to why it does not use technology to scan WhatsApp statuses for indications of likely TVE, WhatsApp repeated the obstacles regarding use of technology on the E2EE parts of its service.

iv. Source of phrases, codes, hashtags

WhatsApp stated that it sourced phrases, codes, and hashtags indicating likely TVE from 'WhatsApp's own ongoing integrity work.'

C. Languages covered by language analysis tools

In response to questions about the languages covered by WhatsApp's language analysis tools, WhatsApp stated that the CT Text Classifier it used to detect new TVE images, videos, and phrases, codes, and hashtags indicating likely TVE is capable of operating in the following languages:

Table I

Arabic	English	French	Hindi	Indonesian	Portuguese
Russian	Spanish	Italian	German		

When asked about the languages covered by Whole Post Integrity Embeddings Service, which WhatsApp also uses to detect new TVE images, videos, and phrases, codes, and hashtags indicating likely TVE on parts of its service, WhatsApp stated that the tool is language agnostic.

In response to follow-up questions from eSafety, WhatsApp stated that the CT Text Classifier and Whole Post Integrity Embeddings Service are language agnostic models trained on text in the following languages:

Table J

Afrikaans	Albanian	Amharic	Arabic	Armenian	Assamese
Azerbaijani	Basque	Belarusian	Bengali	Bengali Romanised	Bosnian
Breton	Bulgarian	Burmese	Catalan	Chinese (Simplified)	Chinese (Traditional)
Croatian	Czech	Danish	Dutch	English	Esperanto
Estonian	Filipino	Finnish	French	Galician	Georgian
German	Greek	Gujarati	Hausa	Hebrew	Hindi
Hindi Romanised	Hungarian	Icelandic	Indonesian	Irish	Italian
Japanese	Javanese	Kannada	Kazakh	Khmer	Korean
Kurdish (Kurmanji)	Kyrgyz	Lao	Latin	Latvian	Lithuanian
Macedonian	Malagasy	Malay	Malayalam	Marathi	Mongolian
Nepali	Norwegian	Oriya	Oromo	Pashto	Persian

Polish	Portuguese	Punjabi	Romanian	Russian	Sanskrit
Scottish Gaelic	Serbian	Sindhi	Sinhala	Slovak	Slovenian
Somali	Spanish	Sundanese	Swahili	Swedish	Tamil
Tamil Romanised	Telegu	Telegu Romanised	Thai	Turkish	Ukrainian
Urdu	Urdu Romanised	Uyghur	Uzbek	Vietnamese	Welsh
Western Frisian	Xhosa	Yiddish			

D. Action taken on TVE

In response to questions about what action was taken when known or unknown TVE images, video, or written material was detected by its tools, WhatsApp stated that 'the signal is used for prioritising content for human review.'

E. Livestreamed TVE

i. Detecting livestreamed TVE

The Notice specified that livestreaming includes one-on-one video calls and video calls where one or more multiple people stream material to a group of any size.

In response to questions about the measures WhatsApp had in place to detect the livestreaming of TVE on its service, WhatsApp provided the following information:

Table K

Parts of service	Measures in place to detect TVE in livestreams?	Interventions used	Names of tools used
Video calls	No	N/A	N/A

In response to why it did not have any measures in place to detect livestreamed TVE in video calls, WhatsApp stated

> While WhatsApp does provide a video calling feature, it does not provide a broadcast live streaming feature. Video calls are limited to 32 participants. Video calls on WhatsApp are end-to-end encrypted, which means that WhatsApp technically cannot proactively monitor the contents of a video call. However, users are able to report other users on a video call.

ii. Reducing the likelihood of livestreamed TVE

In response to questions about the steps taken by WhatsApp to reduce the likelihood that TVE could occur in livestreams, WhatsApp stated that it used the following measures:

- Restrictions for those who have previously violated terms of service or community guidelines/standards – including preventing groups that have been suspended for violating TVE policies from accessing group calling features.
- A 32-participant limit on the number of participants in a video call.

F. Blocking links to TVE material

i. Detection and sources of URLs

WhatsApp was asked about its use of lists or databases to proactively detect and block URLs linking to TVE on other platforms. Specifically, WhatsApp was asked about:

- Known URLs linking to websites/services operated by individuals/organisations dedicated to the creation, promotion, or dissemination of TVE or other TVE-related activities
- URLs linking to known TVE material on other services/websites (which may not be dedicated to TVE)
- Join-links to groups, Channels, communities, or forums on other services that were known to be associated with TVE.

Table L

Service	Blocked URLs to websites/services dedicated to TVE?	Blocked URLs linking to known TVE material on other services/websites?	Blocked join-links to groups/channels on other services known to be associated with TVE?	URL sources
WhatsApp	No	No	No	N/A

In response to why URLs to TVE material are not blocked and whether alternative steps were taken to block URLs, WhatsApp stated that it is 'technically unable to use technology to block URLs on [end-to-end] encrypted parts of the service.' WhatsApp stated that it is 'examining the potential value' of blocking URLs in Channels. WhatsApp also stated that it uses classifiers to detect potential TVE in text in all parts of the service identified at Table H.

G. Off-platform monitoring

WhatsApp was asked if it used off-platform monitoring,[157] either provided internally or by third-party services, to identify accounts or Channels on WhatsApp that are dedicated to TVE. WhatsApp stated that it does and that relevant third-party monitoring vendors are engaged by Meta and if a vender identifies any relevant activity on WhatsApp, Meta will escalate the relevant URLs to WhatsApp for investigation.

WhatsApp also stated, '[t]hese third party vendors vary depending on Meta's operational needs. However, they generally include industry experts in online extremism, militant extremism, and foreign terrorist organizations.'

H. Percentage of TVE sent for human review

WhatsApp was asked to provide the percentage of TVE reports it sent for human review and the criteria and thresholds used to determine when reports were sent for human review. WhatsApp stated that because it does not require end-users to select a specific reporting category when reporting TVE, it could not 'determine the number of user reports where the user intended to report TVE specifically.'

As an alternative to this information, WhatsApp provided the number of accounts that were banned or against which other enforcement actions were taken for TVE-related violations and which also had a user report over the last 30 days.

Table M

	Percentage of user reports of TVE sent for human review	Criteria and thresholds used to determine when a user report is sent for human review	Percentage of TVE detected through automated tools sent for human review	Criteria and thresholds used to determine when a report of TVE detected through automated tools is sent for human review
WhatsApp	100%**	'[H]igh level' of confidence the content violates TVE policies. Violations of TVE policies are sent for human review as they require assessment of context.*	100%	'[H]igh level' of confidence the content violates TVE policies. Violations of TVE policies are sent for human review as they require assessment of context.*

[157] Monitoring of activity on other services.

*** WhatsApp reported that thresholds that determine whether the reported content merits human review are calibrated on an ongoing basis to ensure that WhatsApp is enforcing consistently and with high precision.**

**** WhatsApp noted that these user report figures relate to user reports by Australian users and cover the period 1 March 2024 to 30 April 2024 due to its data retention policies.**

I. Percentage of TVE detected proactively

WhatsApp was asked what percentage of TVE was detected proactively, compared to TVE reported by users, trusted flaggers or through other channels for the following parts of its service:

Table N

Service	Percentage of TVE detected proactively	Percentage of TVE reported by users, trusted flaggers or other
WhatsApp	91%*	9%**

*** For percentage of TVE 'proactively detected' WhatsApp reported on instances where it did not receive a report against the relevant account in the 30 days prior to enforcement.**

**** For percentage of TVE 'reported by users, trusted flaggers or other' WhatsApp reported on instances where it did receive a report against the relevant account in the 30 days prior to enforcement.**

WhatsApp noted that these figures represent TVE created by Australian users during the report period.

J. Appeals against TVE-related moderation

In response to a question about how many appeals were made by users for accounts banned or content removed for TVE, where WhatsApp was alerted by automated tools or user reports, and how many of those were successful, WhatsApp provided the following information:

Table O

How WhatsApp was alerted to TVE	Number of appeals made for accounts banned for TVE breach	Number of appeals that were successful for accounts banned	Number of appeals made for material removed for TVE breach	Number of appeals that were successful for material removed
Automated tools	20*	11	WhatsApp reported that it does not remove individual pieces of content.	
User reports	0**	0		

***For 'alerted by automated tools' WhatsApp reported the number of appeals against accounts where it did not receive a report against the relevant account in the 30 days prior to the ban.**

**** For 'alerted by user report' WhatsApp reported the number of appeals against accounts where it did receive a report against the relevant account in the 30 days prior to the ban.**

WhatsApp noted that these figures represent appeals made by Australian users and cover the period 1 March 2024 to 30 April 2024 due to its data retention policies.

eSafety notes that where WhatsApp's automated tools banned an account for TVE-breaches and a user made an appeal, over 50% of these appeals were successful, despite WhatsApp reporting that 100% of TVE detected through automated tools is sent for human review. Although the absolute volumes are low and therefore not necessarily representative, eSafety notes that a high proportion of account bans being successfully overturned on appeal may indicate flaws in the human review process. WhatsApp subsequently stated that 9 of the 11 successful appeals during this period 'were the result of bans propagated from Facebook and Instagram, and were therefore not subject to human review by WhatsApp'[158].

K. Performing checks on files to determine if they are 'suspicious'

In the Notice, eSafety referred to the fact that WhatsApp's website states 'WhatsApp automatically performs checks to determine if a file is suspicious'.[159] eSafety also noted that in response to a non-periodic reporting notice given in August 2022, WhatsApp stated that 'we are unable to perform any form of check on any other content (for example in a gif, file, or photo) for suspicious content or malware, unless it is provided to us via a user report.

In response to questions about the kinds of checks it performs on files, given its statement on its website, WhatsApp stated

> WhatsApp automatically performs checks to determine if a file is suspicious, to ensure that the format is supported on WhatsApp and doesn't crash the app on the User's device. WhatsApp checks the structure of files, such as media container formats, but not content. To protect user privacy, these checks take place entirely on the user's device, and because of end-to-end encryption, WhatsApp can't see the content of the messages or files.

[158] WhatsApp subsequently clarified that accounts detected by WhatsApp's tools are sent for human review before a ban can be applied, but when a ban is propagated from a ban on Facebook or Instagram, it will occur automatically without further human review.
[159] WhatsApp, 'About suspicious files', accessed 26 February 2024, URL: https://faq.whatsapp.com/667552568038157/?cms_platform=iphone&helpref=platform_switcher

6. Questions about resources, expertise, and human moderation

A. Trust and Safety

i. Trust and Safety and other staff

WhatsApp was asked to provide the number of staff that were employed or contracted by WhatsApp to carry out certain functions at the end of the report period. WhatsApp reported that it did not have data available for the date specified in the notice (29 February 2024), instead it provided the following information for 31 December 2023:

Table P

Category of staff	Number of staff
Engineers employed by WhatsApp focussed on trust and safety	117
Content moderators employed by WhatsApp	0*
Content moderators contracted by WhatsApp	1,365
Trust and safety staff employed by WhatsApp (other than engineers and content moderators)	266**

***WhatsApp stated there are 'Nil' content moderators employed by WhatsApp, and that '[c]ontent reviewers are generally employed by Meta's vendors'. WhatsApp further stated that there were 'around 208 employees' focused on WhatsApp in WhatsApp/Meta's global operations team, which focuses on 'work related to review of content (e.g. quality reviews, building protocols, managing contractors etc.).'**

**** WhatsApp reported that this cohort included employees 'working in global operations and other non-engineering tech functions (i.e., product managers, researchers, designers, etc.).'**

WhatsApp stated that these figures represent teams 'who are focused on core trust and safety work' and that they 'do not represent the full spectrum of people working on trust and safety at Meta/WhatsApp.'

ii. Trust and Safety dedicated to minimising TVE

In response to a question asking if WhatsApp had a dedicated trust and safety team responsible for minimising TVE on WhatsApp, WhatsApp answered 'yes', reporting that it has a 'core cross-functional team dedicated specifically to this area of harm'. WhatsApp stated. 'This team's mandate is to identify, and enforce against, Groups, Channels, Communities and 1:1 messages that violate WhatsApp's TVE policies.'

WhatsApp provided the following information about the composition of its team:

Table Q

Name of role/area of expertise	Number of staff	Number of contractors
Product manager	1	N/A
Engineer	3	N/A
Operations project manager	2	N/A

iii. Surge teams to respond to a TVE crisis

WhatsApp was asked if it had a surge team(s) to respond to TVE crises, such as a livestreamed attacked with content disseminated on the service. WhatsApp answered 'yes' and stated that it has '24/7 escalation coverage to respond to crises, including terrorist attacks.' WhatsApp reported that the relevant on-call employees are able to alert policy, operations, and legal teams, and that the size of the surge team will depend on the nature of the event.

B. Languages human moderators operate across

In response to a question about the languages that its human moderators operate across (both employees and contractors), WhatsApp provided the following:

Table R

Languages covered by employees (all languages)	Languages covered by contractors (all languages)	
N/A*	• Arabic • English • Farsi	• Spanish • Urdu • Pashto

***WhatsApp stated that it 'does not track or require any specific language capabilities for trust and safety employees' and 'relies on the language capabilities of its human review teams who are contractors.'**

WhatsApp subsequently stated:

> WhatsApp provides its reviewers with translation tools to enable them to review material in languages other than their native languages.

eSafety notes that the top 5 languages, other than English, spoken in Australian homes are Arabic, Cantonese, Mandarin, Vietnamese and Punjabi.[160] WhatsApp's human moderators do not cover Cantonese, Mandarin, Vietnamese or Punjabi.

C. Median time to reach an outcome to a user report of TVE

WhatsApp was asked to provide the median time taken to reach an outcome[161] after receiving a user report about TVE for the following parts of its service:

Table S

Parts of the service	Reports from users globally*	Reports from users in Australia*
Direct messages (including groups)	25.3 hours	24.13 hours^
Communities	24.8 hours	N/A**
Channels	24.5 hours	25.3 hours^^

*** WhatsApp reported that these figures reflect enforcement action taken against accounts that were banned for TVE-related violations and had also received a user report over the past 30 days. WhatsApp stated that due to the absence of issue-specific reporting options, WhatsApp cannot identify user reports where the user intended to report TVE specifically. WhatsApp also stated that because it does not log enforcement actions against specific user reports, it was 'not possible … to calculate the median time taken to reach an outcome after receiving a user report of TVE with precision.'**

As an alternative metric, WhatsApp provided the median time from when a user report was enqueued for human review due to a potential TVE violation to when an enforcement action was taken. WhatsApp stated that 24 hours is the maximum amount of time between a user report being made and the user report being enqueued for human review. WhatsApp's responses as listed in Table S reflect the assumed maximum 24 hours that any given report spends waiting to be enqueued, plus the median time taken for enforcement action of each category of user report.

^ WhatsApp reported that it stores data related to Australian users for rolling 90-day periods. The information relating to reports from Australian users is limited to the period 9 February 2024 – 8 May 2024 and relates to a total of 4 user reports.

**** WhatsApp stated that it did not receive any reports about TVE in WhatsApp Communities from Australian users between 9 February 2024 – 8 May 2024.**

---

[160] Australian Bureau of Statistics, 'Cultural diversity: Census', 28 June 2021, URL: https://www.abs.gov.au/statistics/people/people-and-communities/cultural-diversity-census/latest-release#:~:text=Top%205%20languages%20used%20at,Punjabi%20(0.9%20per%20cent).

[161] Defined in the Notice as a calculation from 'the time that a user report is made, to a content moderation outcome or decision, such as removing the content, banning the account, or deciding that no action should be taken.'

^^ WhatsApp reported that it stores data related to Australian users for rolling 90-day periods. The information relating to reports from Australian users is limited to the period 9 February 2024 – 8 May 2024 and relates to a total of 4 users.

D. Volunteer moderation

WhatsApp provided the following information in response to questions about the process its volunteer 'Community admins' follow, and the processes WhatsApp has in place to monitor their conduct and uphold moderation standards:

Table T

Question	Details provided by WhatsApp
Did WhatsApp have a standards policy, or similar, outlining the responsibilities and expectations of volunteer admins?	No WhatsApp stated that the responsibility for enforcing its policies 'remains with WhatsApp', and that volunteer 'Community admins' are encouraged to report potentially violative behaviour or content to WhatsApp for review and enforcement 'like all WhatsApp users'.
What training and/or guidance was provided to volunteer Community admins regarding proactive minimisation of TVE and removal of accounts that share TVE.	WhatsApp reported that it provides a dedicated site[162] for Community admins 'to understand their role, the expectations that their community members may have, and the tools at their disposal'. WhatsApp stated that this site, 'includes guidance on establishing and enforcing a specific Community's rules, if the admin chooses to establish such rules. There is no requirement for them to do so, and WhatsApp does not delegate enforcement of its Terms of Service or general policies to Community admins.'
Were users able to make in service reports about volunteer admins in instances where they were failing to meet any required responsibilities and expectations?	Yes* *WhatsApp stated that end-users are able to report a Community via in-service reporting tools. WhatsApp qualified that this does not necessarily allow reporting of the Community admin personally.
If volunteer admins removed an account from a WhatsApp Community for TVE-breaches, were trust and safety staff informed?	No WhatsApp stated that Community admins should report TVE-related violations to WhatsApp.
If WhatsApp's Trust and Safety staff banned a user for a TVE-related violation in a Community, were the volunteer Community admins of that Community notified?	No In response to a question about the alternative steps WhatsApp took to ensure that volunteer admins were alert to the potential increased risk of TVE in a group, WhatsApp stated: > While Community admins can have an important role in setting the expectations and norms for their Communities, WhatsApp does not expect them to monitor for violations of WhatsApp's TVE policies.

[162] WhatsApp, 'Welcome to the Communities Learning Center', URL: https://www.whatsapp.com/communities/learning/. URL supplied by WhatsApp, URL supplied by WhatsApp.

7. Questions about steps to prevent recidivism

A. Measures and indicators

In response to a question about the measures WhatsApp takes to prevent recidivism for TVE-related breaches on its service, **WhatsApp listed a minimal[163] number of indicators that it used to detect users that have previously been banned for TVE breaches.** eSafety has chosen not to publish these indicators to prevent the information being misused.

WhatsApp stated that it used all indicators by default in circumstances where an account was banned to prevent recidivism by that user.

B. Preventing banned group, channel, communities from being recreated

In response to a question about the measures WhatsApp took to prevent banned TVE Groups, Channels or Communities from being recreated, WhatsApp reported that, 'if WhatsApp suspends a Group, Channel or Community belonging to a policy violating organisation, WhatsApp also bans the admin(s) of the Group, Channel or Community.'

C. Applying TVE-related bans to associated accounts

WhatsApp was asked, when it took action against an account for a TVE-related breach, whether it applied bans to associated accounts. eSafety defined 'associated accounts' as 'other users who are associated with the banned user'. WhatsApp reported that in certain contexts, it will apply account strikes as a form of graduated enforcement against accounts associated with a TVE-related breach.

D. Sharing of banned account details with other entities

WhatsApp was asked if it shared details of accounts banned for TVE with the following entities:

[163] eSafety uses the following terms to give an impression of the extent of the indicators used by services in the table below, rather than publishing the specific indicators which could be misused:
- Minimal: A small number
- Several: A moderate number
- Multiple: A significant number.

Table U

Entity	Shared details of accounts banned for TVE?	Details provided by WhatsApp
Facebook	No	N/A
Instagram	No	N/A
Other service providers (Non-Meta)	No	N/A
Law enforcement	Yes	WhatsApp reported that it 'may share such details if there is an imminent threat to life.'
Regulatory or other public authorities	No	N/A
Global Internet Forum to Counter Terrorism	No	N/A
Civil society groups	No	N/A

Reddit Summary

Overview

Reddit Inc was asked about its Reddit service.

Part 1. Questions in relation to terrorism and violent extremism (TVE)

1. Questions about Reddit's definitions of 'terrorist material and activity' and 'violent extremist material and activity'

A. Terrorist material and activity

In response to a question about how Reddit defines 'terrorist material and activity' or a different but equivalent term for the purposes of its terms of service and community guidelines, Reddit referred to its Content Policy (https://www.redditinc.com/policies/content-policy) and responded that it

> prohibits content that glorifies, incites or calls for violence or physical harm, including content that "promotes or supports the activities of terrorists or designated terrorist organizations.

Reddit defined terrorist content as

Violative content includes: propaganda material posted by terrorists or designated terrorist organizations and their supporters, expressions of affiliation or support for terrorists or designated terrorist organizations, and glorification of terrorist acts. It also includes content that solicits or incites a person or group to participate, commit, or contribute to terrorist activities.

B. Violent extremist material and activity

In response to a question about how Reddit defines 'violent extremist material and activity' or an equivalent term for the purposes of its terms of service and community guidelines, Reddit

reported that its Terms and Content Policy does not define 'violent extremist material and activity' but that it more broadly

> prohibits content that glorifies, incites or calls for violence or physical harm (Rule 1), which includes (but is not limited to): credible threats of violence against an individual or group of people; posts containing mass killer manifestos or imagery of their violence; terrorist content, including propaganda; posts containing imagery or text that incites, glorifies, or encourages self-harm or suicide; posts that request, or give instructions on, ways to self-harm or commit suicide; and graphic violence, images, or videos without appropriate context.[164]

Reddit stated that Rule 1 in Reddit's Content Policy more broadly prohibits

> "communities and users that incite violence or that promote hate based on identity or vulnerability," including race, colour, religion, national origin, ethnicity, immigration status, gender, gender identity, sexual orientation, pregnancy, or disability. It also includes victims of a major violent event and their families[165]

2. Thresholds/criteria to determine action on TVE breaches

Reddit was asked if it had criteria or thresholds in place to determine what action would be taken when TVE was identified on its service. Reddit provided the following information:

Table A

Actions taken on accounts or content when TVE was identified	Criteria/thresholds reported
Permanent account ban	Reddit stated that accounts confirmed to have posted terrorist content are permanently banned. Reddit added that in determining the appropriate enforcement action for other TVE-related offences of its violence policy it considers the type and severity of the violation, as well as the user's violation history. Reddit added that egregious or repeated offences will result in a permanent ban of the account.
Temporary suspension	Reddit stated that users may receive a 3-day or 7-day suspension, depending on the account's history and the severity of the violation.

[164] Reddit pointed to a Help Centre article which it said explains Reddit's rule against violent content and violent threats: Reddit, 'Do not post violent content', accessed 26 July 2024, URL: https://support.reddithelp.com/hc/en-us/articles/360043513151-Do-not-post-violent-content. URL supplied by Reddit.

[165] Reddit pointed to a Help Centre article that explains Rule 1 in more detail: Reddit, 'Promoting hate based on identity or vulnerability, accessed 26 July 2024, URL: https://support.reddithelp.com/hc/en-us/articles/360045715951-Promoting-Hate-Based-on-Identity-or-Vulnerability. URL supplied by Reddit.

Account strikes	Reddit reported that warnings and account strikes are reserved for first time and low severity violations of its violence policy, providing the example of a user who may inadvertently violate Reddit's policies while attempting to share content related to newsworthy global events.

3. Questions about reporting of TVE

A. In-service reporting of TVE on different parts of the Reddit service

In response to questions about whether users could report instances of TVE to Reddit within the service (as opposed to navigating to a separate webform or email address), Reddit responded:

Table B

Parts of the service	Accessing Reddit via a browser	Accessing Reddit via an app
	In-service reporting option	
Subreddits	Yes	Yes
Chat	Yes	Yes
Private messages	Yes	Yes
Channels	Yes	Yes
Subreddit Wikis	No	No

Reddit reported that for all in-service reporting of TVE, whether via a browser or an app, users can choose the reporting category 'Threatening violence' to report TVE.

In relation to reporting content in subreddit wikis, Reddit responded that subreddit wikis are optional resource pages controlled by community moderator teams and which users who are not mods of the associated subreddit do not control and cannot post to by default. Reddit noted that, as mod-controlled resource pages, subreddit wikis may therefore be reported through the Moderator Code of Conduct Violation report form. Reddit reported that it is

> in the process of implementing the ability for users to report subreddits from the subreddit page, which will take into account subreddit wikis.

B. User reporting of TVE when not signed in

In response to questions about whether users could report instances of TVE with specific reporting categories to Reddit when not signed in, Reddit responded:

Table C

	Able to report TVE through the following applications
Access via web browser	No
Access via a Reddit app	Yes for iOS No for Android

Reddit stated that reports from a logged-in state help with prioritisation and that 'frivolous reports' are a lower priority than reports from users with 'a history of accuracy in reporting'.

Reddit reported that users not signed in who are accessing Reddit via a web browser or via the native Reddit Android app can report content via a web form on Reddit's Help Centre.[166]

C. Reporting of TVE by third party services that use Reddit's API

In answer to a question about whether Reddit has minimum safety requirements for third party services that use Reddit's APIs[167] to access its service, Reddit responded that it does have minimum safety requirements and that this includes the requirement for user reporting functions on third party apps to notify Reddit of breaches of its terms of service.[168] Reddit provided a link to its Developer Terms[169] which it said outline how third parties may use Reddit's services, including its Data API and Reddit data including user-generated content. Reddit also pointed to its Data API Terms[170] which 'obligate third parties that have their own website, webpage, application, bot, service, research, or other offering (an "App") that allows end users to submit or provide content to the App to have appropriate notice and takedown processes and to comply with all applicable laws'.

Reddit reported that third party apps are not required to provide user reporting categories specific to TVE.

---

[166] Reddit referred to the web form on its Help Centre: Reddit, 'Submit a request', accessed 26 July 2024, URL: https://support.reddithelp.com/hc/en-us/requests/new?ticket_form_id=15968767746196. URL supplied by Reddit.

[167] Application programming interface. Reddit's approach to use of APIs was updated recently, see: Reddit, 'Creating a healthy ecosystem for Reddit data and Reddit data API access', 18 April 2023, accessed 26 February 2024, URL: https://www.redditinc.com/blog/2023apiupdates.

[168] Reddit referred to section 7.4 of its Developer Terms: Reddit, 'Developer Terms', last revised 4 March 2024, accessed 26 July 2024, https://www.redditinc.com/policies/developer-terms. URL supplied by Reddit.

[169] Reddit referred to section 3.5 its Developer Terms: Reddit, 'Developer Terms', last revised 4 March 2024, accessed 26 June 2024, https://www.redditinc.com/policies/developer-terms. URL supplied by Reddit.

[170] Reddit referred to its Data API Terms: Reddit, 'Data API Terms, last revised 18 April 2023, accessed 26 July 2024, URL: https://www.redditinc.com/policies/data-api-terms. URL supplied by Reddit.

Reddit stated that third parties accessing Reddit data are obliged to remove any content they accessed via Reddit's developer services that was subsequently deleted by Reddit users or Reddit. Reddit added that this includes any content that Reddit removed for violating its Content Policy.[171] Reddit also stated that it provides instructions and automated means to third parties to make data deletion easy.[172]

D. Reporting mechanisms for other entities to report TVE

In answer to a question about having separate reporting mechanisms for other entities to report TVE, Reddit responded that it does have reporting mechanisms (separate from users in general) for law enforcement[173]; trusted flaggers[174] and regulatory or other public authorities [175][176]

Reddit stated that reporting by these entities via dedicated reporting channels ensured that, 'requests from law enforcement, government authorities, and flaggers with expertise in identifying illegal content are routed directly to the team with expertise to handle such requests'.

Reddit reported that it does not have a separate reporting mechanism for

- Civil society groups

Reddit stated that it does receive alerts from specific civil society organisations such as Tech Against Terrorism's TCAP alerts and that civil society groups and other entities can report via the report form on Reddit's Help Centre or the standard reporting tool for logged-in users.

---

[171] See definition provided by Reddit on page 1.
[172] Reddit referred to section 3.5 of its Developer Terms: Reddit, 'Developer Terms', last revised 4 March 2024, accessed 26 June 2024, https://www.redditinc.com/policies/developer-terms. URL supplied by Reddit.
[173] Reddit referred to its Guidelines for Law Enforcement: Reddit, 'Guidelines for law enforcement', last revised 13 March 2024, accessed 26 July 2024, URL: https://www.redditinc.com/policies/guidelines-for-law-enforcement. URL supplied by Reddit.
[174] Reddit referred to its EU illegal content report form for people claiming legal rights in the EU and trusted flaggers designated under the EU Digital Services Act (DSA) to submit reports of terrorist content: Reddit, 'EU illegal content report form', accessed 26 July 2024, URL: https://support.reddithelp.com/hc/en-us/requests/new?ticket_form_id=19623931614484. URL supplied by Reddit.
[175] Reddit referred to the email address legalcontentreview@reddit.com for law enforcement, regulatory, and other public authorities to submit content removal or review requests.
[176] Reddit referred to the email address LETCO@reddit.com for designated EU authorities to submit removal orders relating to terrorist content pursuant to the EU's Terrorist Content Online Regulation (Regulation (EU) 2021/784) (TCOR).

4. Questions about proactive detection

A. Detecting known material using hash-matching

i. Known TVE images

In response to questions about hash-matching for known TVE images, Reddit provided the following information:

Table D

Parts of service	Used image hash-matching tools?	Names of tools used
Subreddit (public) **Subreddits (private)**	Yes	• Snooron – Internal hash-matching functionality • Rule-Executor-V2 (REV2) – automated enforcement system
Chat **Channels**	No – but since implemented	Implemented since reporting period: • Snooron – Internal image hash-matching functionality • Rule-Executor-V2 (REV2) – automated enforcement system
Account profile picture **Subreddit profile picture** **Channel profile picture**	No	Reddit stated it is 'currently building new internal hash tooling which will supplement detection' in these parts of its service.
Private messages	N/A	Reddit reported that 'images cannot be sent via pm'
Subreddit Wikis	N/A	Reddit reported that it 'does not support image upload directly to wikis'

In response to why hash-matching tools were not used on chat and chat channels, Reddit stated that it had 'prioritised integration into parts of the service where video was shared', but that it was 'currently in the process of integrating its relatively new terrorism hash set into chat and chat channels' and that it was also 'currently building new internal hash tooling which will supplement detection efforts in chat and chat channels'.

eSafety notes that since its response to the Notice, Reddit updated eSafety that it had completed implementation of detection via its existing internal hash sets into chat and chat channels.

In response to why hash-matching tools were not used on account profile pictures, subreddit profile pictures, and channel profile pictures, Reddit stated 'as indicated above [referring to

chat and chat channels], Reddit is currently building new internal hash tooling which will be utilised to enable detection via TVE hashes on account, subreddit, and channel profile photos'.

In response to what alternative reasonable steps Reddit was taking to detect known TVE images on chat, channels, account profile pictures, subreddit profile pictures, and channel profile pictures, Reddit responded that it was currently building new internal hash tooling to enable detection of TVE hashes on profile photos, subreddit profile photos, and channel profile photos. Reddit also noted that it uses third-party tooling that leverages machine learning to predict the likelihood that any given media asset (e.g. image or video) contains terrorist content (e.g. via the presence of watermarks or logos), and that it uses various detection methods, including both automated detection and user reports, to detect TVE content posted by accounts across the site, including within subreddits, chat, and channels.

eSafety notes that Reddit is not a current GIFCT member which, combined with the GIFCT's policy change, means that Reddit does not have access to the GIFCT's current hash database.

ii. Known TVE video

In response to questions about hash-matching for known TVE video, Reddit provided the following information:

Table E

Parts of service	Used video hash-matching tools?	Names of tools used
Subreddits (public)	Yes	• Snooron – Internal hash-matching functionality • Rule-Executor-V2 (REV2) – automated enforcement system
Subreddits (private)		
Chat	N/A	Reddit reported that 'video may not be sent via chat'
Private messages	N/A	Reddit reported that 'video may not be sent via private message

iii. Known TVE written material

In response to questions about hash-matching for known TVE written material, such as manifestos or text promoting, inciting, instructing TVE, Reddit provided the following information:

Table F

Parts of service	Used written material hash-matching tools?	Names of tools used
Subreddits (public) **Subreddits (private)**	Yes	• Snooron – Internal image hash-matching functionality • Rule-Executor-V2 (REV2) – automated enforcement system
Chat **Channels**	No (but since implemented)	Implemented since reporting period: • Snooron – Internal image hash-matching functionality • Rule-Executor-V2 (REV2) – automated enforcement system
Private messages	N/A	Reddit reported that 'images/screenshots may not be sent via private message'
Subreddit Wikis	N/A	Reddit reported that 'Reddit does not support image/screenshot upload directly to wikis'

Since its response to the Notice, Reddit updated eSafety that it had completed implementation of detection via its existing internal hash sets into chat and chat channels.

iv. Sources of TVE hashes

Reddit reported that it sourced its hashes of known TVE images, video and written material from the following databases:

- Reddit's own TVE hash list from various sources,* including content on Reddit confirmed to be terrorist content.
- Reddit stated that it intends to take all the hashes from the Tech Against Terrorism TCAP Archive, and that it is currently trialling the process.

Reddit noted that it retrieved all hashes from the GIFCT hash-sharing database until late 2022 and that when it was retrieving those hashes it would do so every 5 minutes.

Reddit also noted that document files such as word documents or PDFs are not hashed as Reddit does not allow the upload of these types of files to its platform.

*Reddit noted that it's threat detection team sourced screenshots, images and videos to hash and add to its hash depository from a variety of sources, including:

- In-house experts
- Content moderation specialists
- The intelligence community
- US National Counterterrorism Center (NCTC) Liaison Office
- Expert NGOs

- Industry partners
- Tech Against Terrorism's (TAT) Terrorist Content Analytics Platform (TCAP) – Reddit reported that it is also working with TAT on TAT's new hash bank facility.
- GIFCT's Hash Sharing Consortium – Reddit reported that it was shut off from this program in September 2022 when the GIFCT restricted access to members only. Reddit said that it continues to use the hashes that it received (up until September 2022) to identify potential terrorist content on its platform.

v. Action taken on known TVE

In response to questions about what action was taken when known TVE images, video or written material were detected by its tools, Reddit responded that

- images, videos and written material that have not already been confirmed to include terrorist content are sent for human review; or
- if the content has already been confirmed as terrorist content, it is removed via automation.

B. Detecting new TVE material

i. New or 'unknown' TVE images

In response to questions about detection of new (or 'previously unknown') TVE images, Reddit provided the following information:

Table G

Parts of service	Used tools for images?	Names of tools used
Subreddits (public)	Yes	• Hive AI - AI image detection tooling; image optical character recognition (OCR) • Rule-Executor-V2 (REV2) – automated enforcement system
Subreddits (private)		
Chat		
Channels		
Account profile pictures		
Channel profile pictures		
Subreddit profile pictures	Yes	• Hive AI - AI text detection tooling
Private messages	N/A	Reddit reported that 'It is not possible to share video or images via private message'
Subreddit Wikis	N/A	Reddit reported that it 'does not support image upload directly to wikis'

ii. New or 'unknown' TVE videos

In response to questions about detection of new (or 'previously unknown') TVE video, Reddit provided the following information:

Table H

Parts of service	Used tools for video?	Names of tools used	Whether tools are video and/or audio classifiers, or others
Subreddits (public)	Yes	• Hive AI – video classification AI • Rule-Executor-V2 (REV2) – automated enforcement system • Google Vision OCR API – text detection	Video and text classifiers
Subreddits (private)			
Chat	N/A	Reddit reported that 'videos may not be sent via chat'	N/A
Private messages	N/A	Reddit reported that 'videos may not be sent via private message'	N/A

When asked to specify whether the tools used to detect new TVE videos are video and/or audio classifiers Reddit responded that they are video and text classifiers.

When asked what languages the technology used to detect new TVE videos Reddit responded that

> its text classifiers and automated enforcement system can detect new TVE videos based on the text included in those video posts (e.g., the post title).

And that, 'Our threat detection team may create detection rules in any language, depending on the needs of the incident/event at hand. Our text classifier tooling will identify content in the language of the rule as entered. Our third-party AI video detection tooling is configured for English language analysis'.

In response to follow-up questions from eSafety, Reddit clarified that as at 29 February 2024, its tools for detecting new TVE videos, and phrases, codes, and hashtags indicating likely CSEA operate in the same languages as those used to detect likely TVE material (see **Tables J, K and L**).

iii. Action taken on new TVE images and videos

In response to questions about what action was taken when Reddit detected new TVE images or videos, Reddit stated that:

- potential new TVE content that has been detected by Reddit's own automated enforcement system or third-party AI detection tools is sent for human review
- if content confirmed as terrorist content, it is removed from the platform
- the account that posted the content is permanently banned.

Reddit added that if new TVE content is detected by Reddit's text classifiers (including image OCR) and automated enforcement system it

- automatically removes the content from the platform
- 'Users may also receive an account level sanction as appropriate for the behaviour, which may include a permanent ban on the account'.

Reddit also added that 'new terrorist/TVE media is hashed' to prevent future sharing.

iv. Text Analysis to detect TVE

In response to questions about technology used to detect phrases, codes, hashtags indicating likely TVE in text (for example manifestos or text promoting, inciting, instructing (TVE), Reddit provided the following information:

Table I

Parts of service	Used text analysis tools?	Names of tools used
Subreddits (public)	Yes	• Snooron – Keyword matching text classifier functionality • Rule-Executor-V2 (REV2) – automated enforcement system • Hive AI - image optical character recognition (OCR)
Subreddits (private)		
Chat		
Channels		
Private messages	Yes	• Snooron – Keyword matching text classifier functionality • Rule-Executor-V2 (REV2) – automated enforcement system
Account name		
Account profile description		
Subreddit name		

Subreddit profile description		
Channel name	No	
Channel profile description		
Subreddit Wikis		

In response to why technology to detect phrases and codes is not used on channel name and description, Reddit responded that chat channels are a relatively new product for Reddit and that it was still integrating these into their text classifier and automated enforcement system.

In response to why technology to detect phrases and codes is not used on subreddit wikis, Reddit responded that

> Subreddit wiki pages are not intended as a place for users to share content but for volunteer community moderators to post and organise information related to their subreddits

and that Reddit

> have not observed patterns of abuse of subreddit wikis for the purpose of sharing harmful content, and…the vast majority of subreddits have disabled this feature.

Reddit noted that its automated tools use text classifiers and machine learning to detect TVE content in chat channels and subreddits, and its third-party AI detection tooling detects potential terrorist images in channel profile pictures.

Reddit noted that language analysis is integral to its efforts to addressing TVE on its platform.

v. Source of phrases, codes, hashtags

Reddit reported that its threat detection team sourced its lists of indicators from a wide range of sources as per the list outlined under ‘Known TVE images’ above.

vi. Action taken on likely written TVE

In response to a question about what action was taken when these indicators were detected by its tools, Reddit responded that

- phrases, codes or keywords indicating likely TVE are automatically removed from the platform
- media confirmed as TVE is hashed to prevent future sharing
- ‘Users may also receive an account level sanction as appropriate for the behaviour, which may include a permanent ban on the account’.

When asked if Reddit blocks words or phrases that it detects indicating likely TVE to users searching for them, Reddit responded that it 'does not currently block users from searching for words or phrases indicating likely TVE because such words and phrases are highly entwined with legitimate searches for news and other information about important world affairs'. Reddit added that instead it focusses its efforts on the various human and automated measures used to prevent likely TVE from appearing on its platform thus avoiding unnecessary constraints on users who are following its rules.

C. Languages covered by language analysis tools

When asked what languages the technology used to detect phrases, codes and hashtags indicating likely TVE in text Reddit responded that it does not have a hashtag functionality and that its threat detection team 'may create detection rules in any language, depending on the needs of the incident/event at hand' and 'Our text classifier tooling will identify content in the language of the rule as entered'.

In response to follow-up questions from eSafety, Reddit clarified that as at 29 February 2024, it uses a keyword matching text classifier function of its internal tool Snooron, to detect known TVE images and videos, and phrases, codes, and hashtags indicating likely TVE in text. Reddit reported that Snooron operates in the following languages:

Table J

Arabic	Bengali	Cantonese	Dutch	English	French
German	Hebrew	Hindi	Indonesian	Italian	Japanese
Mandarin	Portuguese	Romanian	Russian	Spanish	Spanish MX
Turkish	Ukrainian	Norwegian	Danish	Finnish	Swedish
Vietnamese	Slovak				

Reddit noted that its automated enforcement system operates in the same languages as its text classification tool.

Reddit also reported that it uses an Optical Character Recognition (OCR) tool, which utilises Hive AI to detect new TVE images, and phrases, codes, and hashtags indicating likely TVE in text. Reddit reported that this Hive AI OCR tool is capable of recognising text in the following languages:

Table K

English	Spanish	French	German	Italian	Mandarin
Russian	Portuguese	Arabic	Korean	Japanese	Hindi

Reddit reported that it uses Google Vision OCR API to detect text in new TVE videos. Reddit provided a link to the languages supported by Google Vision OCR API tool:[177]

Table L

Afrikaans	Albanian	Arabic	Armenian	Belarusian	Bengali
Bulgarian	Catalan	Chinese	Croatian	Czech	Danish
Dutch	English	Estonian	Filipino	Finnish	French
German	Greek	Gujarati	Hebrew	Hindi	Hungarian
Icelandic	Indonesian	Italian	Japanese	Kannada	Khmer
Korean	Lao	Latvian	Lithuanian	Macedonian	Malay
Malayalam	Marathi	Nepali	Norwegian	Persian	Polish
Portuguese	Punjabi	Romanian	Russian	Serbian	Slovak
Slovenian	Spanish	Swedish	Tagalog	Tamil	Telugu
Thai	Turkish	Ukrainian	Vietnamese	Yiddish	

Reddit also stated that it is currently developing an internal tool. Once implemented, Reddit stated that this tool will support 80 languages.

D. Blocking links to TVE material

i. Detection and sources of URLs

Reddit was asked about its use of lists or databases to proactively detect and block URLs linking to TVE on other platforms. Specifically, Reddit was asked about:

- Known URLs linking to websites/services operated by individuals/organisations dedicated to the creation, promotion, or dissemination of TVE or other TVE-related activities
- URLs linking to known TVE material on other services/websites (which may not be dedicated to TVE)
- Join-links to groups/channels on other services that were known to be associated with TVE

[177] Reddit provided the following link to the list of languages supported by Google Vision OCR API URL: Google, 'OCR language support', accessed 26 June 2024, URL: https://cloud.google.com/vision/docs/languages. URL supplied by Reddit.

Table M

Parts of service	Used databases/lists of known URLs to block URLs to websites/services?	Blocked URLs linking to known TVE material on other services/websites?	Blocked join-links to groups/channels on other services known to be associated with TVE?	URL sources
Subreddits (public)	Yes	Yes	Yes	Reddit reported that its threat detection team sourced URLs/domains from the various sources as per the list outlined under 'Known TVE images' above. Including: • Research on third-party websites or forums • Information shared by third-parties
Subreddits (private)	Yes	Yes	Yes	
Chat	Yes	Yes	Yes	
Private messages	Yes	Yes	Yes	
Channels	Yes	Yes	Yes	
Account profile description	Yes	Yes	Yes	
Subreddit profile description	Yes	Yes	Yes	
Channel profile description	No	No	No	
Subreddit Wikis	No	No	No	

In response to why URLs are not blocked on channel profile description, Reddit responded that 'channel profile descriptions is text only.' and that 'Unlike account and subreddit profiles, social links may not be added to channel descriptions'.

In response to why URLs are not blocked on subreddit wikis, Reddit responded, as per response under 'Text Analysis' above.

Reddit also noted it uses various detection methods to detect TVE content posted in chat channels and in subreddits, including text and media classifiers, ML detection models, and use reports.

ii. Action taken on accounts attempting to share blocked URLs/join-links

In response to a question about what action was taken when an account was detected attempting to share a blocked URL dedicated to TVE, a blocked URL linking to known TVE on another website/service or a blocked join-links to groups/channels on other services known to be associated with TVE, Reddit responded that Reddit's tools block submissions of banned

domains to the platform and that 'Posts or other content containing banned links cannot be submitted'.

E. Off-platform monitoring

In response to a question about whether Reddit used off-platform monitoring[178], either provided internally or by third-party services, to identify accounts, subreddits or channels present on its service dedicated to TVE, Reddit responded that 'off-platform monitoring is an integral part of Reddit's threat detection efforts and allows Reddit to proactively identify new threats, actors, tactics, and TVE material to hash'.

Reddit reported that its threat detection team undertake a number of monitoring activities that eSafety has chosen not to publish to prevent this information being misused.

Following a subsequent question from eSafety, Reddit reported that it is part of a 'multi-party contractual partnership intended to enable the sharing of information on threat activity between participating industry partners'.

F. Percentage of reports sent for human review

In response to questions about the percentage of TVE reports sent for human review and the criteria and thresholds used to determine when reports are sent for human review, Reddit provided the following information:

Table N

	Percentage of user reports of TVE sent for human review	Criteria and thresholds used to determine when a user report is sent for human review	Percentage of TVE detected through automated tools sent for human review	Criteria and thresholds used to determine when a report of TVE detected through automated tools is sent for human review
Reddit	100%*	• 'Possible propaganda material of a designated foreign terrorist organisation' • Specific indicators of terrorist organisation affiliation	66.5%**	• Tool 90% and above confidence of terrorist content • Hash match of terrorist content not previously confirmed by Reddit human moderators***

[178] Monitoring of activity on other services.

		• 'Content that solicits or incites a person/group to • participate, commit, or contribute to terrorist activities.' • 'First-person or real time/on the ground media of terrorist violence (with reasonable exceptions for citizen journalism or other newsworthy content).'		

*** Reddit reported that the 100% refers to reports that users have made under its 'threatening violence' option and that Reddit has thereafter determined may be terrorist content.**

**** Reddit reported that the 66.5% refers to 'terrorist content' (as opposed to 'TVE') detected through automated tools that is sent for human review.**

***** Reddit reported that a hash match of a media asset such as image or video that has previously been confirmed as terrorist content by a Reddit human moderator will automatically be removed.**

G. Percentage of TVE detected proactively

Reddit was asked what percentage of TVE was detected proactively, compared to TVE reported by users, trusted flaggers or through other channels for the following parts of its service:

Table O

Parts of the service	Percentage of TVE* detected proactively	Percentage of TVE* reported by users, trusted flaggers or through other channels
Subreddits (public)	79.4%	20.6%
Subreddits (private)	100%	0%
Chat	Reddit reported that during the report period it did not have any terrorism-related removals in these parts of the service	
Private messages		
Channels		
Subreddit Wikis		

*** Reddit stated that when it actions content under its 'violence policy' it categorises those removals either under the 'broader violence category' or the 'narrower terrorism sub-subcategory' not as 'TVE'.**

Reddit noted that a single item of content may be flagged in multiple ways given there are a number of tools operating at the same time to identify violating content and for this reason Reddit categorised content by how it was first reported, either via report or via proactive detection.

H. Appeals against TVE-related moderation

In response to a question about how many appeals have been made by users for accounts banned or content removed for TVE, where the service was alerted by automated tools or user reports, and how many of those were successful, Reddit provided the following information:

Table P

How Reddit was alerted to TVE	Number of appeals made for accounts banned for TVE breach*	Number of appeals that were successful for accounts banned*	Number of appeals made for material removed for TVE breach*	Number of appeals that were successful for material removed*
Automated tools	29	0	Reddit reported that it does not currently have this data**	
User reports	92	2		

*** Reddit stated that when it actions content under its 'violence policy' it categorises those removals either under the 'broader violence category' or the 'narrower terrorism sub-subcategory' not as 'TVE'.**

**** Reddit reported that it was unable to provide appeals volumes for material removed due to a TVE breach, explaining that its appeals process, during the report period, was linked to account-level sanctions and not to content-level sanctions. Reddit said it is 'in the process of building the capacity to provide such breakdowns going forward.'**

5. Questions about resources, expertise and human moderation

A. Trust and Safety

i. Trust and Safety and other staff

Reddit was asked to provide the number of staff employed or contracted by Reddit to carry out certain functions as at 29 February 2024. Reddit provided the following information:

Table Q

Category of staff	Number of employees	Number of contractors	Total
Engineers employed by Reddit focussed on trust and safety	82	7	89
Content moderators	15	107	122
Trust and safety staff employed by Reddit (other than engineers and content moderators)	71	23	94

Reddit noted that as of 29 February 2024, the total number of Reddit employees was 2030 and the total number of Reddit contractors was 989.

Reddit also noted that

> Reddit's various safety teams consist of a diverse range of roles, functions, and subject matter expertise, including content moderation, engineering, threat analysis, data science, research, training, trust & safety policy, legal and community policy & enforcement.

ii. Trust and Safety dedicated to minimising TVE

In response to a question asking if Reddit had a dedicated trust and safety team(s) responsible for minimising TVE on Reddit, Reddit responded that it has 'multiple teams' and referred to:

- Its threat detection team which identifies off-platform risks and sources material to inform its various detection tools, and which manages systems to utilise this information.
- A dedicated team with expertise in reviewing content flagged as potential terrorist content following automated detection or a user report.

Reddit provided the following information about the composition of these teams:

Table R

Name of role/area of expertise	Number of staff	Number of contractors
Trust & Safety Policy	2	0
Safety Operations	24	120

iii. Surge teams to respond to a TVE crisis

Reddit was asked if it had a surge team(s) to respond to TVE crises such as a livestreamed attack with content disseminated on the service, Reddit responded that it does but clarified that it does not have a livestreaming function.

Reddit provided the following information about the composition of this team:

Table S

Name of role/area of expertise	Number of staff	Number of contractors
Trust & Safety Policy	4	1
Safety Operations	27	0*
Community	3	0
Public Policy	1	0

*** Reddit reported that its content moderation contractors are not part of its official incident response teams, but that they are notified of incidents and provided with special guidance as appropriate)**

Reddit noted that it has a dedicated incident response protocol, and outlined the following:

- The protocol is governed under its broader Trust and Safety teams;
- Sets out processes and responsibilities around response to incidents such as livestreamed attacks on other services;
- Establishes a dedicated internal communications channel for the given incident to ensure cross-functional visibility and coordination on all actions;
- An incident 'commander' leads an incident response team made up of personnel from Table S as well as policy, legal, communications and community enforcement specialists.

Reddit added that it's response to terrorist incidents, such as the 7 October attack in Israel, involve

> a dedicated surge team of trained individuals with linguistic and subject-matter expertise who can assist in our review queues and in outreach to our volunteer community moderators should we see an increase in violative content (either as a result of user reports or automated detection efforts).

B. Languages human moderators operate across

In response to a question about the languages human moderators, both employees and contractors, operated across, Reddit responded:

Table T

Languages covered by employees (all languages)	Languages covered by contractors (all languages)
• English • French • Spanish • Portuguese • Arabic • Russian • German • Turkish • Urdu • Hindi • Telugu • Shona • Zulu	• English • French • Spanish • Portuguese • Russian • Turkish • Hindi • German

eSafety notes that the top 5 languages, other than English, spoken in Australian homes are Arabic, Cantonese, Mandarin, Vietnamese and Punjabi.[179] Reddit's human moderators do not cover Cantonese, Mandarin, Vietnamese or Punjabi.

C. Median time to reach an outcome to a user report of TVE

Reddit was asked to provide the median time taken to reach an outcome[180] after receiving a user report about TVE for the following parts of its service:

Table U

Parts of the service	Reports from users globally	Reports from users in Australia
Subreddits (public)	62.2 hours*	31.3 hours*
Subreddits (private)	Reddit reported that there were no user reports that Reddit confirmed to be terrorist content on these parts of its service during the report period	
Chat		
Private messages		
Channels		
Subreddit Wikis		

[179] Australian Bureau of Statistics, 'Cultural diversity: Census', 28 June 2021, URL: https://www.abs.gov.au/statistics/people/people-and-communities/cultural-diversity-census/latest-release#:~:text=Top%205%20languages%20used%20at,Punjabi%20(0.9%20per%20cent).

[180] Defined in the Notice as a calculation from 'the time that a user report is made, to a content moderation outcome or decision, such as removing the content, banning the account, or deciding that no action should be taken.'

*** Reddit noted that users may report material that may be terrorist and/or violent extremist material under the violence reporting option, or potentially under the hate reporting option. Reddit further noted that it has no way to distinguish a user report of TVE from non-TVE violations of these rules, and that it therefore does not have data on the median time taken to reach an outcome after receiving "user reports of TVE" on the service. Reddit also noted that reports that its human safety team determines may relate to terrorist content are sent to a specialised terrorism queue for further human review. Reddit initially provided the time taken to respond to a user report from the time 'a ticket was escalated to our terrorism review queue', which was 2.2 hours for users globally and 1 hour for users from Australia. Following a subsequent question to Reddit, it provided the median time between user report and ticket closure for reports escalated to Reddit's specialised terrorism queue.**

Reddit added that

> unlike content that is flagged through Reddit's automated terrorist content detection efforts, user reports which Reddit determines may relate to terrorist content go through a two-step review process to ensure that the content is reviewed by a subject matter expert.

Reddit also highlighted that 'only 19.8% of content removed as terrorist content during the reporting period was first flagged to us by a user report, and only 16% of the content escalated to specialist review was first flagged by a user report.' It noted that it considered the response time of its terrorism specialists was therefore 'the most accurate picture'.

D. Volunteer moderation

Reddit provided the following information in response to questions about the processes its volunteer moderators follow, and the processes Reddit has in place to monitor their conduct and uphold moderation standards:

Table V

Question	Response
Did Reddit have a standards policy, or similar, outlining the responsibilities and expectations of volunteer moderators?	Yes. Reddit stated that 'Moderators are expected to abide by the Reddit Moderator Code of Conduct[181]...which sets out Reddit's expectations for community moderators – including the expectation that mods uphold Reddit's Content Policy, in addition to making a concerted effort to remove and report violating content in their communities'.
What training and/or guidance was provided to Reddit volunteer moderators regarding proactive minimisation of TVE and	Reddit pointed to:

[181] Reddit provided a link to its Moderator Code of Conduct: Reddit, 'Moderator Code of Conduct', effective 3 July 2024, accessed 26 July 2024, URL: https://www.redditinc.com/policies/moderator-code-of-conduct. URL supplied by Reddit.

removal of accounts that share TVE.	The training and guidance materials provided for volunteer community moderators in its Moderator Help Centre[182], specifically outlining: • Content Policy • Moderator Code of Conduct ○ Including chapter on crisis management In response to global events: • 'Community Relations team frequently reaches out to moderators of potentially impacted communities to share situational guidance on a bespoke basis' Proactive reminders re: • Availability of automated content control tools (automoderator) • Moderator Reserves program[183]
Were users able to make in service reports about volunteer moderators in instances where they were failing to meet any required responsibilities and expectations?	No. Reddit reported that users may report violations of the Moderator Code of Conduct using a form on the Help Centre.
If volunteer moderators removed an account from subreddits and/or channels (both public and private) for TVE-breaches, were trust and safety staff informed?	Reddit responded 'Yes' that trust and safety staff are informed when a volunteer moderator removes an account from subreddits and/or channels (both public and private) for TVE breaches. Reddit reported that user reports of policy breaches go to both the moderation of teams of the subreddit where the content was posted and to Reddit and therefore that Reddit will already be aware of any content removed by a volunteer moderator as a result of a user report. Following a subsequent question from eSafety, Reddit reported that it is not automatically informed when a volunteer moderator removes an account from a subreddit or chat channel. Reddit stated that the ban and reason (if the volunteer moderator chooses to record one) will be visible to the Reddit staff when they review the removed account – along with all other subreddit bans enacted against the account by volunteer moderators. Reddit also stated that any policy breaches proactively found by volunteer moderators or reported as breaching their specific subreddit rules can result in the volunteer moderator removing the content and removing the user from their community. Reddit stated that the volunteer moderators, as per other users, are encouraged to report violating content to Reddit.

[182] Reddit provided a link to its Moderator Help Centre: Reddit, 'Moderator Help', accessed 26 July 2024, URL: https://support.reddithelp.com/hc/en-us/p/mod_help_center

[183] Reddit reported that this program 'allows existing mod teams to draw from a team of vetted supplemental volunteer moderators in the event of temporary, abnormal surges in traffic.'

If Reddit's Trust and Safety staff banned a user for a TVE-related violation in a subreddit or channel, were the volunteer moderators of that subreddit or channel notified?	Reddit responded 'sometimes'.
	
	To ensure volunteer moderators were alert to an increased risk of TVE in a subreddit or channel Reddit reported that:
	
	In response to global events:
	• Provide situational guidance on violent and terrorist content policies and how they should be enforced at community level.
	
	In response to 'uptick in users posting violating violent content in a particular subreddit':
	• Community team alert moderators of subreddit to the trend and ensure they understand policies and to remind moderators about tools and programs that Reddit offers to assist in managing communities in times of crisis, including automated content filters and the Moderator Reserves program.
	  ○ Thresholds for engaging moderators are dependent on:
	  ○ Ongoing global incidents
	  ○ Nature of incidents
	  ○ Observed behaviours of subreddit the moderator team
	
	In response to upticks in violative content due to ongoing issues in subreddits, including lack of active moderation:
	• Community team engage with moderator teams.
	• Restrictive measures may be imposed
	  ○ Removing moderators
	  ○ Ensure moderator teams approve all posts one at a time prior to being visible by community
	  ○ Banning community from platform
	
	In response to subreddits appearing to be dedicated to posting violative content or if it has no moderators:
	• Subreddit removed entirely with no outreach from community team

In response to a question about the action taken by trust and safety staff when they became aware of volunteer moderator decisions relating to TVE, such as removing a user from a subreddit or channel, Reddit reported that the following process is taken:

- Automated tools help prioritise user reports.
- If a human reviewer determines that content may include terrorist content it is flagged and routed to a specialist.

- If it is determined that content breaches the violent content policy, including the policy against terrorist content, the content is removed from the platform and action taken against the user who posted.
- Appropriate enforcement action depends on type and severity of violation, including users' violation history. Examples of enforcement action:
    - Permanent account ban
    - Initial warning, then 3-day ban, then 7-day ban, then permanent ban

eSafety notes that Reddit's response above is in relation to receipt of a policy breach report from any user – not from a volunteer moderator specifically.

6. Questions about steps to prevent recidivism

A. Measures and indicators

Reddit reported that it had measures in place to prevent recidivism for TVE-related breaches on its service and provided the following information:

For egregious TVE-related offences:

- Account may be permanently banned
- Account holder may be banned

For less egregious offences, defined by Reddit as cases where users inadvertently violate policies while sharing content related to newsworthy global events:

- User education on how and why they have violated policies
- Account may be permanently banned – but account holder may not be banned from creating new accounts
- Account holder permanently banned if violates policies multiple times using multiple accounts – new accounts will also be banned

Ban evasions:

- Users may report suspected ban-evading subreddits via forms in Help Centre
- Reddit reported other safeguards and procedures against ban evasion that eSafety has chosen not to publish to prevent this information being misused.

Subreddit-level ban evasion:

- Ban Evasion Filter[184] available for moderators to use.

  Reddit reported that this tool 'filters the participation of accounts' that are related to accounts recently banned by moderators on their subreddits. Reddit added that the signals that this tool picks up from Reddit's backend system are not revealed to moderators for privacy reasons.

Reddit listed multiple[185] indicators to detect users that have previously been banned for TVE-related breaches and provided additional indicators that it will be incorporating, which eSafety has chosen not to publish to prevent the information from being misused.

Reddit stated that it used all indicators by default in all instances where an account was banned to prevent recidivism by that user.

B. Prevention of subreddit and channel recreation following ban

Reddit reported that it has ban evasion detection tooling to prevent subreddits and channels from being recreated after they have been banned. Reddit provided information on the measures and several indicators it has in place which eSafety has chosen not to publish to prevent the information from being misused.

Reddit also noted that should a subreddit evade its ban evasion detection efforts, the various detection methods outlined in this summary, which detect violative content and flag subreddits with high volumes of violative content for review, would ensure that any subreddits created for the purpose of evading prior subreddit bans would be uncovered.

C. Accounts associated with accounts banned for TVE-related breaches

Reddit reported that it did apply bans (or other action) to accounts associated with an account banned for TVE-related breaches. (Associated accounts could be members of the same TVE-related subreddits or channels).

Reddit provided the following criteria/thresholds for taking action on associated accounts:

- Subreddits dedicated to sharing content that violates policies may result in a ban of the entire subreddit and moderation team.

---

[184] Reddit provided a link to information about its Ban Evasion Filter, Reddit, 'Ban Evasion Filter', accessed 26 July 2024, URL: https://support.reddithelp.com/hc/en-us/articles/15484544471444-Ban-Evasion-Filter. URL supplied by Reddit.

[185] eSafety uses the following terms to give an impression of the extent of the indicators used by services in the table below, rather than publishing the specific indicators which could be misused:
- Minimal: A small number
- Several: A moderate number
- Multiple: A significant number.

- Indicators that flag new or alternative accounts connected to banned accounts may be banned.
- Individuals who violate Reddit policies multiple times, across multiple accounts, are permanently banned from the platform and any new or alternative accounts detected are also banned.

D. Sharing of banned account details with other entities

Reddit was asked if it shared details of accounts banned for TVE with the following entities:

Table W

Entity	Shared details of accounts banned for TVE?	Details provided by Reddit
Other service providers	Yes	Reddit stated that it 'has information sharing agreements in place with many other platforms, intended to enable sharing of information related to potential threats.'
Law enforcement	Yes	Reddit stated that it 'proactively reports imminent threats to life or safety to law enforcement, including TVE-related threats.'
Regulatory or other public authorities	No	
Global Internet Forum to Counter Terrorism	No	
Civil society groups	No	

7. Questions about Reddit recommender systems

A. Preventing amplification of TVE

i. Recommender algorithm – interventions

In answer to a question about whether Reddit had interventions in place to prevent the amplification of TVE via its recommender systems, Reddit provided the following information:

- Reddit removes TVE when it is identified on the service – including through its use of the proactive detection tools summarised at Section 4.
    - 'The various tools referred to in our response to this reporting notice – including use of hash technology, proprietary and third party ML models, and keyword detection – help Reddit to prevent the amplification of TVE content via its

recommendation systems. When we identify TVE content, we remove the content from the platform.'

- Reddit periodically rates communities based on the content within those communities using an internal taxonomy rating system:
    - 'Content from communities with certain ratings, such as violence, are not eligible for recommendation.'
    - Content from unrated communities is not eligible for recommendation.
    - Communities must meet certain size and activity thresholds to be eligible for rating … brand new communities spun up in response to a particular event, or those without a strong record of constructive behaviour, cannot be amplified.
- Content must achieve a suitability score to be eligible for recommendation surfaces, like home feed suggestions.
    - Criteria impacting this score change and are constantly updated, but include things like downvotes, user reports, machine learning content analysis, and other safety signals.
- Reddit's subreddit structure limits virality
    - Interest-based subreddit structure means that content of interest to one community may not be of interest to another.
    - Specific subreddit rules, such as a text-only rule or a 'must be about cats' rule limits sharing across subreddits and will likely violate subreddit rules or be downvoted.

ii. Recommender algorithm – testing

In answer to a question about any testing Reddit performs to ensure that its recommender systems do not amplify TVE, Reddit provided the following information:

- Model Experimentation – 'product and machine learning teams use a metric to minimise a user's interaction with policy violating content (assessed by content that is later lagged and removed). If we observe a significant increase in interaction with policy violating content, an investigation is initiated to resolve the overshoot.'
- Model design, development, deployment, monitoring/feedback – Impact assessment carried out at each stage.

Reddit added

> For high impact models, an internal model cards process outlines assumptions, relevant model factors, and safety considerations with a cross functional group of stakeholders. Evaluations focus on mitigating key risks, which include but are not limited to bias

(particularly the disenfranchisement of protected groups), exposure to sensitive content, toxic behaviour, and policy breaking content.

iii. Recommender algorithm – positive interventions

In response to questions about having systems in place to stage positive interventions, for example by promoting deradicalizing content for at risk users when a user seeks out TVE material, or if certain phrases or keywords linked to TVE are blocked for users seeking that content, Reddit responded that it does not have these measures in place. Reddit reiterated that it 'does not currently block users from searching for words or phrases indicating likely TVE because such words and phrases are highly entwined with legitimate searches for news and other information about important world affairs'. Reddit added that instead it focusses its efforts on the various human and automated measures used to prevent likely TVE from appearing on its platform thus avoiding unnecessary constraints on users who are following its rules.

iv. Voter algorithm

In response to a question about the measures Reddit had in place to ensure that its voter algorithm was prevented from amplifying illegal and harmful content such as TVE, Reddit responded that, as opposed to other platforms where positive or negative content can result in heightened visibility, the ability of Reddit users to upvote or downvote content will result in the rise or fall of that content based on those votes. Reddit stated

> Rule-violating, inaccurate, suspicious, or simply disrespectful posts or comments are often downvoted to oblivion, limiting their visibility in the individual subreddit where the content was posted and also on Reddit's post aggregation feeds.

Reddit also reported that it uses vote manipulation detection models to prevent attempts to game the voting system. Accounts detected trying to game the voting system may face account-level sanctions and votes may be thrown out, depending on the behaviour observed.

Reddit added that its upvote/downvote system works in tandem with its "karma" feature. Reddit explained that karma is a publicly visible reputation score which is based on the number of upvotes and downvotes received on an account's posts and comments. Reddit stated that users will take this score into account when deciding whether to trust content posted by that account and that users are therefore encouraged to make valuable, interesting, insightful and positive contributions in their interactions to ensure a higher karma score.

Reddit also added that regardless of how highly upvoted content may be, Reddit's internal classifications for communities and content will determine what content can be included in feeds.

Part 2. Questions in relation to child sexual exploitation and abuse (CSEA)

8. Questions about reporting of CSEA

A. In-service reporting of CSEA on different parts of the Reddit service

In response to questions about whether users could report instances of CSEA to Reddit within the service (as opposed to navigating to a separate webform or email address), Reddit responded:

Table X

Parts of the service	Accessing Reddit via a browser	Accessing Reddit via an app
	In-service reporting option	
Subreddits	Yes	Yes
Chat	Yes	Yes
Private messages	Yes	Yes
Channels	Yes	Yes
Subreddit Wikis	No	No

Reddit reported that for all in-service reporting of CSEA, whether via a browser or an app, users can choose the reporting category 'Minor abuse or sexualisation' and can then select from three options (i) sexual or aggressive content; (ii) predatory or inappropriate behaviour; (iii) content involving physical or emotional abuse or neglect.

In relation to reporting content in subreddit wikis, Reddit responded, in the same way it responded to the same question regarding TVE, that users do not control subreddit wikis, but rather subreddit wikis are resource pages controlled by community moderator teams and therefore any reports about community moderators can be made through the Moderator Code of Conduct Violation report form.

9. Questions about proactive detection of CSEA

A. Detecting known material using hash matching

i. Known CSEA images

In response to questions about hash matching for known CSEA images, Reddit provided the following information:

Table Y

Parts of the service	Used image hash matching tools?	Names of tools used
Subreddits (public)	Yes	PhotoDNA
Subreddits (private)	Yes	PhotoDNA
Chat	Yes	PhotoDNA
Channels	Yes	PhotoDNA
Account profile pictures	Yes	PhotoDNA
Subreddit profile pictures	Yes	PhotoDNA
Channel profile pictures	Yes	PhotoDNA
Private messages	N/A	Reddit reported that 'images may not be shared via private message'
Subreddit Wikis	N/A	Reddit reported that it 'does not support image upload directly to wikis'

Reddit reported that it takes a subset of hashes from the following hash databases:

- NCMEC –hashes from the 'NGO' database
- Microsoft PhotoDNA - NCMEC hash set, Cybertip.ca (CCA), and Canadian Technology Industry (CIH)

Reddit also stated that it is currently building its own child sexual abuse material (CSAM) hash set.

With regards to how often Reddit updates its hashes of CSEA images Reddit responded:

- NCMEC hashes – daily
- Microsoft PhotoDNA hashes - Since mid-2022, Microsoft provided Reddit with the PhotoDNA technology but not direct access to the hash database(s). Reddit use a local copy of the PhotoDNA technology to detect potential CSEA images, and then call Microsoft's PhotoDNA

cloud API to confirm potential matches. Database updates are managed on the Microsoft side so Reddit reported that it was always working with their currently active hash set.

ii. Known CSEA video

In response to questions about hash matching for known CSEA video, Reddit provided the following information:

Table Z

Parts of service	Used video hash matching tools?	Names of tools used
Subreddits (public)	Yes	YouTube CSAI Match
Subreddits (private)	Yes	YouTube CSAI Match
Chat	N/A	Reddit reported that 'video may not be shared via chat'
Private messages	N/A	Reddit reported that video may not be shared via private message

Reddit reported that it takes all hashes from YouTube CSAI Match.

Reddit also stated that it is currently building its own CSAM hash set for video.

With regards to how often Reddit updates its hashes of CSEA videos Reddit responded that it 'currently calls YouTube's CSAI API for every video uploaded to Reddit; we do not maintain a local hash set. Database updates are managed on the provider side (i.e., YouTube), so we are always working with their currently active hash set'.

B. Detecting new CSEA material

i. New or 'unknown' CSEA images

In response to questions about detection of new (or 'previously unknown') CSEA images, Reddit provided the following information:

Table AA

Parts of service	Used tools to detect new CSEA images?	Names of tools used
Subreddits (public)	Yes	• Hive AI – image optical character recognition (OCR) • Rule-Executor-V2 (REV2) – automated enforcement system
Subreddits (private)		
Chat		
Channels		
Account profile pictures		

Channel profile pictures		
Subreddit profile pictures	No	
Private messages	N/A	Reddit reported that 'no images or video may be sent via private message'
Subreddit Wikis	N/A	Reddit reported that it 'does not support image upload directly to wikis'

In response to why tools were not used to detect new CSEA images on subreddit profile pictures, Reddit responded that its tools to detect CSEA posted to subreddits or in account profiles and the subsequent removal of new communities dedicated to CSEA are 'most effective'. Reddit also added that its subreddit ban evasion detection tools help to prevent subreddits and channels from being recreated after they have been banned.

ii. New or 'unknown' CSEA video

In response to questions about detection of new (or 'previously unknown') CSEA video, Reddit provided the following information:

Table BB

Parts of service	Used tools to detect new CSEA video?	Names of tools used	Whether tools are video and/or audio classifiers, or other
Subreddits (public)	Yes	• Rule-Executor-V2 (REV2) – automated enforcement system • Google Vision OCR API	Text classifiers
Subreddits (private)			
Chat	N/A	Reddit reported that 'videos may not be sent via chat'	N/A
Private messages	N/A	Reddit reported that 'videos may not be sent via private message'	N/A

eSafety notes that although Reddit is using tools to detect new CSEA images and video these tools do so based on the text included in the image, video and video posts (e.g. the post title) and not through other indicators in the image or video (e.g. nudity detection and age estimation). This may mean key indicators of CSEA are missed.

When asked to specify whether the tools used to detect new CSEA videos are video and/or audio classifiers Reddit responded that they are text classifiers. When asked what languages the technology used to detect new CSEA videos Reddit responded that

> its text classifiers and automated enforcement system can detect new CSEA videos based on the text included in those video posts (e.g., the post title).

And Reddit's 'threat detection team may create rules in any language, and our text classifier tooling will identify content in the language of the rule as entered'.

In response to follow-up questions from eSafety, Reddit clarified that as at 29 February 2024, its tools for detecting new CSEA videos operate in the same languages as those used to detect likely TVE material (see **Table J, K and L**).

C. Action taken on known and new CSEA

In response to questions about what action was taken when known and new CSEA images and videos, and known terms, abbreviations and codes were detected by its tools Reddit responded that:

- The content is blocked/removed from Reddit and the account is permanently banned
- An enforcement ticket is created and prioritised for human review
- Depending on the outcome of human review, Reddit may make a report to NCMEC and take further enforcement action (including account sanctions).

D. Text Analysis of CSEA

In response to questions about language analysis technology used to detect terms, abbreviations, codes and hashtags indicating likely CSEA in particular but not limited to grooming, sexual extortion and the trading and sale of CSEA material on various parts of its service, Reddit provided the following information:

Table CC

Parts of service	Used text analysis tools to detect likely CSEA?	Names of tools used
Subreddits (public)	Yes	• Snooron - Keyword matching text classifier • Rule-Executor-V2 (REV2) – automated enforcement system • Hive AI - image optical character recognition (OCR)
Subreddits (private)		
Chat		
Channels		
Private messages	Yes	• Snooron - Keyword matching text classifier • Rule-Executor-V2 (REV2) – automated enforcement system
Account profile description	Yes	• Snooron - Keyword matching text classifier • Rule-Executor-V2 (REV2) – automated enforcement system
Subreddit profile description	Yes	
Account name	Yes	
Subreddit name	Yes	
Channel profile description	No	N/A
Channel name	No	N/A
Subreddit Wikis	No	N/A

In response to why language analysis technology to detect terms, abbreviations and codes is not used on channel name and description, Reddit responded, as it did to the same question regarding TVE, that chat channels are a relatively new product for Reddit and

> full integration of channel names and descriptions into our text classifier tooling and automated enforcement system is currently in progress.

In response to why technology to detect phrases and codes is not used on subreddit wikis, Reddit responded, as it did to the same question regarding TVE, that

> Subreddit wiki pages are not intended as a place for users to share content but for volunteer community moderators to post and organise information related to their subreddits

and that Reddit

> have not observed patterns of abuse of subreddit wikis for the purpose of sharing harmful content, and…the vast majority of subreddits have disabled this feature.

i. Sources of terms, abbreviations and codes

Reddit reported that its 'threat detection team source CSEA-related keywords and indicators from a wide range of sources', including:

- NCMEC
- Expert NGOs
- In-house experts – Reddit's threat detection team and trust and safety policy team conduct research on Reddit and on other platforms to 'identify new trends and indicators, including scaled spamming efforts targeting multiple platforms'.
- Industry partners – share and receive information to inform and improve detection and enforcement efforts and thus stop the spread of harmful content.

ii. Languages covered by language analysis tools

When asked what languages the technology used to detect terms, abbreviations, codes and hashtags indicating likely CSEA, Reddit responded that it does not have a hashtag functionality and that its threat detection team 'may create detection rules in any language, and our text classifier tooling will identify content in the language of the rule as entered'.

In response to follow-up questions from eSafety, Reddit clarified that as at 29 February 2024, its tools for detecting phrases, codes, and hashtags indicating likely CSEA operate in the same languages as those used to detect likely TVE material (see **Table J and K**).

E. Blocking links to CSEA material

i. URLs linking to known CSEA

In response to a question about whether Reddit blocked URLs linking to known CSEA, Reddit provided the following information:

Table DD

Parts of service	Used databases/lists of known URLs to block URLs to websites/services?	URL sources
Subreddits (public)	Yes	Reddit reported that its threat detection team proactively sourced CSEA-related indicators, including domains, from a range of sources, including the list outlined under 'Text analysis' – 'Sources of phrases, codes, hashtags' above.
Subreddits (private)	Yes	
Chat	Yes	
Private messages	Yes	
Channels	Yes	
Account profile description	Yes	
Subreddit profile description	Yes	

Channel profile description	No	N/A
Subreddit Wikis	No	N/A

In response to a question about what action was taken when an account was detected attempting to share a blocked URL to known CSEA, Reddit responded that its tools block submissions of banned domains to the platform and that 'Posts or other content containing banned links cannot be submitted'.

In response to why URLs are not blocked on channel profile description and subreddit wikis, Reddit responded, as it did to the same question regarding TVE, that 'channel profile descriptions is text only' and that 'Unlike account and subreddit profiles, social links may not be added to channel descriptions' and that it has 'not observed patterns of abuse of subreddit wikis for the purpose of sharing harmful content'.

Reddit also added that it is

> migrating our domain ban tooling to a new system and is working on plans to expand it to cover wikis.

F. Percentage of CSEA detected proactively

Reddit was asked what percentage of CSEA was detected proactively, compared to CSEA reported by users, trusted flaggers or through other channels for the following services:

Table EE

Parts of the service	Percentage of CSEA detected proactively	Percentage of CSEA reported by users, trusted flaggers or through other channels
Subreddits (public)	34.10%	65.9%
Subreddits (private)	35.22%	64.78%
Chat	90.40%	9.60%
Private messages	49.45%	50.55%
Channels	17.83%	82.17%
Subreddit Wikis	Reddit reported that during the report period it did not have any CSEA-related removals in subreddit wikis	

Reddit noted that a single item of content may be flagged in multiple ways given there are a number of tools operating at the same time to identify violating content and for this reason Reddit categorised content by how it was first reported, either via report or via proactive detection.

eSafety notes that there is considerable variation in CSEA detection rates between proactive detection as compared to content reported by users, trusted flaggers, others across Reddit's services. >90% of CSEA is proactively detected on Chat. Conversely >80% of CSEA is reported by users, trusted flaggers or others on Channels, even though the same automated tools are used on both Chat and Channels and the same reporting categories to report CSEA are offered to users.

G. Appeals against CSEA-related moderation

In response to a question about how many appeals have been made by users for accounts banned or content removed for CSEA, where the service was alerted by automated tools or user reports, and how many of those were successful, Reddit provided the following information:

Table FF

How Reddit was alerted to CSEA	Number of appeals made for accounts banned for CSEA breach	Number of appeals that were successful for accounts banned	Number of appeals made for material removed for CSEA breach	Number of appeals that were successful for material removed
Automated tools	3,766	89	Reddit reported that it does not currently have this data*	
User reports	4,076	159		

*** Reddit reported that it was unable to provide appeals volumes for material removed due to a CSEA breach, explaining that its appeals process, during the report period, was linked to account-level sanctions and not to content-level sanctions. Reddit said it is 'in the process of building the capacity to provide such breakdowns going forward.'**

10. Questions about resources, expertise and human moderation

A. Median time to reach an outcome to a user report of CSEA

Reddit was asked to provide the median time taken to reach an outcome[186] after receiving a user report about CSEA for the following parts of its service:

[186] Defined in the Notice as a calculation from 'the time that a user report is made, to a content moderation outcome or decision, such as removing the content, banning the account, or deciding that no action should be taken.'

Table GG

Parts of the service	Reports from users globally	Reports from users in Australia
Subreddits (public)	12.9 hrs	12.4 hrs
Subreddits (private)	12.4 hrs	6.8 hrs
Chat	18.6 hrs	17.1 hrs
Private messages	12.9 hrs	12.0 hrs
Channels	24.8 hrs	29.5 hrs
Subreddit Wikis	Reddit reported that there were no CSEA-related Moderator Code of Conduct reports relating to subreddit wikis during the report period	

Reddit reported that it calculated the above metrics from

> the earliest time there was a CSEA-related report on a particular piece of content in the relevant part of the service, per user, and calculated the time between that report and the ultimate decision on that report.

Reddit also reported that it looked at content reported as

> sexual or suggestive content involving a minor or predatory or inappropriate behaviour involving a minor, regardless of the decision outcome or ultimate action reason.

11. Questions about steps to prevent recidivism

A. Measures and indicators

In response to a question asking if Reddit had measures in place to prevent recidivism for CSEA-related breaches on its service, Reddit provided the following information:

For egregious offences

- Permanent suspension of accounts that share CSAM or engage in predatory or inappropriate behaviour towards minors
- Users reported to NCMEC are banned from creating new accounts on Reddit

For less egregious offences, defined by Reddit for example as ‘posting lewd comments on an otherwise acceptable photo’

- Depends on type and severity of violation, and the user’s violation history
    - May first receive a warning, then 3-day ban, 7-day ban, then permanent ban

- Account holder is permanently banned after receiving multiple account level bans

Ban evasions:

- Users may report suspected ban-evading subreddits via forms in Help Centre

  Reddit reported other safeguards and procedures against ban evasion that eSafety has chosen not to publish to prevent this information being misused.

Reddit listed multiple[187] indicators to detect users that have previously been banned for CSEA-related breaches and provided additional indicators that it will be incorporating, which eSafety has chosen not to publish to prevent the information from being misused.

Reddit stated that it used all indicators by default in all instances where an account was banned to prevent recidivism by that user.

12. Additional information

In response to an opportunity to provide further information and context to any of its responses to the questions asked in the Notice, Reddit added that

> Reddit has zero tolerance for content or interactions that involve terrorism or sexual exploitation or abuse of minors. Combating this type of content is a top priority for our safety teams. We enforce our policies strictly across the platform, and are committed to continually evolving and strengthening our methods and tools.
>
> In addition, it's important to note that our rule against the abuse of minors is not limited to CSEA material, but also includes other inappropriate or abusive content and behaviour involving minors, both sexual and non-sexual, including neglect, physical or emotional abuse. For example, videos of things like physical school fights are not allowed on the platform.

[187] eSafety uses the following terms to give an impression of the extent of the indicators used by services in the table below, rather than publishing the specific indicators which could be misused:
- Minimal: A small number
- Several: A moderate number
- Multiple: A significant number.

Telegram summary

Overview

Telegram FZ LLC was asked about its Telegram service.

Part 1. Questions in relation to Terrorism and Violent Extremism (TVE)

1. Questions about Telegram's definitions of 'terrorist material and activity' and 'violent extremist material and activity'

A. Terrorist material and activity

In response to a question about how Telegram defines 'terrorist material and activity' or a different but equivalent term for the purposes of its terms of service and community guidelines, Telegram stated:

> For the purposes of its moderation procedures, Telegram is guided by the notion of "terrorist content" that comprises texts, imagery, recordings, and footage promoting and glorifying violence and terrorist ideology, soliciting funds for terrorist causes, instructing or advising on planning or carrying out of terror attacks.

B. Violent extremist material and activity

In response to a question about how Telegram defines 'violent extremist material and activity' or a different but equivalent term for the purposes of its terms of service and community guidelines, Telegram stated that, in its moderation procedures, it defines 'violent extremist content' as:

> texts, imagery, recordings, and footage advocating for violence against a person or a group (i.e., specific threats of physical harm, etc.), as well [sic] instructions for creating and obtaining weapons, explosives and other means of carrying out violent attacks.

Telegram also stated that in its moderation procedures, it defined 'violent content' as:

> [content which] also covers graphic, gruesome or shocking materials, like graphic details of torture—unless such content is clearly communicated for the purposes of news reporting or raising awareness of human rights violations.

2. Prohibiting 'illicit content' on private parts of Telegram

On 18 March 2024, when Telegram was given the Notice, the Telegram Terms of Service stated:

> by signing up for Telegram, you accept our Privacy Policy and agree not to:
>
> - Use our service to send spam or scam users
> - Promote violence on ***publicly viewable*** [emphasis added] Telegram channels, bots, etc.
> - Post illegal pornographic content on ***publicly viewable*** [emphasis added] Telegram channels, bots, etc.[188]

eSafety highlighted this in the Notice, and asked Telegram to specify whether these rules permitted end-users to promote violence or post illegal pornographic content on private parts of the service – namely, Private Channels, Groups, and Secret Chats.

Telegram stated that this was not the case, and that the 'Telegram Terms of Service … apply throughout the app, regardless of chat type.' Telegram stated that it 'doesn't tolerate illicit content' in these parts of the service, and that it will 'take appropriate action within its technical capabilities whenever it becomes aware of such content'.

Telegram stated that the references quoted by eSafety in Telegram's terms of service to publicly viewable parts of the service refer to the areas of its service that Telegram moderators proactively monitored, but that is not to say that Telegram 'tolerate[d] illicit content in private channels, groups, or secret chats'.

Telegram provided the following information about how it detected violative material in the private parts of Telegram where moderators cannot proactively check messages:

- User reporting – Telegram stated that content in private Communities[189] can be reported by users. Telegram also stated that 'regular users, non-registered viewers, and organisations' can report material using in-service reporting options, or through dedicated email addresses.[190] Telegram stated that messages reported by end-users in private Communities

[188] Telegram, 'Terms of Service', accessed 9 February 2024, URL: https://telegram.org/tos
[189] Telegram used the term 'Communities' to refer generally to groups and channels on the service.
[190] Telegram provided the following e-mail addresses for reporting violative material on the service: abuse@telegram.org and StopCA@telegram.org.

(i.e., Channels and Groups) 'are forwarded to moderators', but messages reported in Secret Chats are not (see section 2B).

- Proactive detection measures – Telegram stated that it used 'algorithmic detection measures that prevent abuse' on its service. Telegram referred to its use of hash-matching to detect known TVE and CSEA material on public and private Communities, matched against previously identified TVE and CSEA on Telegram's public content. Telegram also referred to its use of 'automated content detection and manual search strategies tailored to locate Communities engaged in violations of Telegram Terms of Service'.

eSafety notes that responses captured in sections 5 and 9 highlight that there was inconsistent use of proactive detection tools across the 'private' parts of Telegram's service. Telegram also did not take any external hashes from external organisations which share hashes of terrorism and violent extremism.[191]

eSafety notes that limiting hash matching exclusively to material that Telegram itself has previously seen and removed risks missing TVE material that Telegram has not detected yet, and this material continuing to circulate on the platform even when such material has already been identified by other online service providers and hashed in extensive shared databases like those run by the GIFCT or Tech Against Terrorism.

A. Moderating 'private' Communities using 'invite links'

Telegram stated in circumstances where a 'private' Community is made accessible to the broader public via an 'invite link', it also changes Telegram's ability to monitor that Community. Telegram gave the example that if an 'invite link' to a private Community is shared on a public part of Telegram or another social media service, then

> the Community is considered to be public for content moderation purposes (thanks to the fact that moderators can follow the link and view the messages within before any user reports are made).

Telegram stated that it routinely detects such 'invite links' when the public channels or groups hosting them are taken down, and removes the associated Communities 'if appropriate based on the content they publish'.

[191] Following consultation with Telegram on the proposed report for publication, Telegram reported that it 'routinely reviewed hash databases compiled by Europol to inform its systems for proactive detection.'

B. Moderating E2EE 'Secret Chats'

Telegram added that it has 'no technical means of verifying the accuracy of user reports regarding content stored' inside Secret Chats because these messages are protected by E2EE. Telegram stated that messages in Secret Chats were not 'forwarded' to moderators when they were reported by an end-user.

Without access to the messages being reported, Telegram reported that it relies on alternative signals or indicators to determine if 'the reported user is not otherwise engaging in harmful or malicious behaviour'. eSafety has chosen not to publish these alternative signals to prevent them being misused.

eSafety notes that there are alternative measures that enable content moderators to review E2EE messages that have been reported by end-users as harmful or otherwise violative. For example, WhatsApp (which is E2EE) has processes in place that enable its moderators to receive the last 5 messages sent to an end-user from the account they are reporting.[192] eSafety considers that having measures in place that enable moderators to review the material being reported by end-users is key to ensuring that these reports can be responded to effectively.

C. Changes to Telegram's FAQs in September 2024

On 18 March 2024, when Telegram was given the Notice, Telegram's 'frequently asked questions' web page stated:

> Q: There's illegal content on Telegram. How do I take it down?
>
> All Telegram chats and group chats are private amongst their participants. We do not process any requests related to them.
>
> But sticker sets, channels, and bots on Telegram are publicly available. If you find sticker sets or bots on Telegram that you think are illegal, please ping us at abuse@telegram.org.

eSafety pointed to this in the Notice and asked Telegram to specify the steps it was taking to comply with the Expectations in relation to the safe use of private chats and private group chats, including after a user report was made about illegal or harmful content, given this statement. Telegram responded by referring to its use of proactive moderation tools, user reporting tools, and specialised moderation teams to address abuse on its service. Telegram stated

[192] WhatsApp, 'About reporting and blocking someone on WhatsApp', accessed 15 October 2024, URL: https://faq.whatsapp.com/414631957536067/

> If a report is confirmed, Telegram acts with due regard to all known circumstances. Disseminating terrorist and violent content or CSAM on any parts of Telegram service leads to permanent removal of associated accounts and Communities.

Regarding the statement on its 'frequently asked questions' page that it would not process removal requests for 'illegal content' on certain parts of its service, Telegram stated this information was outdated and was the result of statements about Telegram's stance on copyright infringement having been mistakenly copied to a section dealing with Telegram's stance on illegal content. Specifically, Telegram stated

> As at February 29, 2024, certain portions of the Telegram FAQ may have featured outdated information, some of which was updated in September 2024 (with further updates planned in the coming months). This included the item quoted under "Context" in this question, which mistakenly included text from an earlier revision of the FAQ. Namely, the mention of "not processing requests regarding private chats" had been erroneously copied from the FAQ section related to copyright infringement, where it is present to this day.[193]
>
> The quoted paragraph was aimed at law-abiding Telegram users who rely on Telegram for the privacy for their personal communication. It was meant to emphasize that, as Telegram moderators cannot proactively inspect the private messages of its users, they cannot act on unsupported requests which do not rely on reporting mechanisms.
>
> The text was never meant to imply that Telegram's Terms of Service could be violated in private chats. This is evidenced by the fact that users could always report both incoming private and secret chats, and messages in private groups and channels to moderators.

eSafety notes that online media outlet, The Verge, first reported this change to Telegram's FAQs page as having occurred on 6 September 2024.[194]

eSafety also notes that the Internet Archive's Wayback Machine shows that the item Telegram states was 'mistakenly included text from an earlier version of the FAQ' was present on Telegram's FAQ page as far back as 15 March 2016, and that this item appears to pre-date any reference to copyright infringement in Telegram's FAQs.[195]

---

[193] Telegram, 'Telegram FAQ - Q: A bot or channel is infringing on my copyright. What do I do?', URL supplied by Telegram on 13 September 2024, URL: https://telegram.org/faq#q-abot-or-channel-is-infringing-on-my-copyright-what-do-i-do

[194] The Verge, 'Telegram changes its tone on moderating private chats after CEO's arrest', 6 September 2024, URL: https://www.theverge.com/2024/9/5/24237254/telegram-pavel-durov-arrest-private-chats-moderation-policy-change

[195] Internet Archive Wayback Machine, 'Telegram FAQ – 15 March 2016', accessed 16 October 2024, URL: http://web.archive.org/web/20160315182715/https://telegram.org/faq

3. Thresholds/criteria to determine action on TVE breaches

Telegram was asked if it had criteria or thresholds in place to determine what action would be taken when TVE was identified on Telegram. Telegram provided the following information:

Table A

Actions taken on accounts or content when TVE was identified	Criteria/thresholds reported for Telegram
Permanent account ban	Telegram stated the following: • Disseminating material that calls for violence in the form of text, image, recordings, footage or otherwise. Telegram specified this means material 'like concrete and specified threats of physical harm'. • Disseminating material that is gruesome or shockingly graphic. Telegram gave such examples as 'graphic details of torture, accident photos' or material that 'glorif[ies] or promote[s] violent or terrorist ideologies'. • Soliciting funds for terrorist organisations or causes. • Owning or being an administrator of a Community involved in the above activities.
Account strikes	Telegram stated that if a Community, or an account belonging to a 'journalist' or 'researcher', reposts TVE with the intention of sharing 'legitimate scientific research, historical records, or news', then Telegram may either: • grant an exception; or • apply up to two warnings before terminating the Community or account. Telegram stated the decision on enforcement depends on the 'severity, purpose and relevance of the posted content under applicable law'.

Telegram also stated that 'where appropriate', it will remove publications in Communities and remove associated groups and channels.

4. Questions about reporting of TVE

A. In-service reporting of TVE on different parts of Telegram

In response to questions about whether users could report instances of TVE to Telegram within the service (as opposed to navigating to a separate webform or email address), Telegram responded:

Table B

Parts of the service	In-service reporting option?	Reporting category
Chats	Yes	'Block user > Report Spam'*
Secret Chats	Yes	
Group chats (public)	Yes	'Violence'
Group chats (private)	Yes	
Channels (public)	Yes	
Channels (private)	Yes	
Stories	Yes	
Voice calls	No**	
Video calls	No**	

*In response to a follow up question from eSafety, which highlighted that in eSafety's testing on the Telegram iOS app, for Chats and Secret Chats the option to 'Report spam' was not present in all cases. Telegram subsequently clarified that the 'Block + Report Spam' reporting flow is only available when the Chat or Secret Chat is 'initiated by non-contacts and strangers'. eSafety understands that when an end-user wishes to report a message from an account they have already added as a contact, the only option in-service is to 'Block user'.

Telegram provided the following reason for this discrepancy in reporting functionality

> In the extremely unlikely event that a user's friend or acquaintance began sending them TVE content, Telegram contends that it would be more reasonable and effective for said user to contact authorities directly, providing all relevant proof and contact information.

eSafety considers that limiting reporting tools to scenarios where the account sending harmful or violative material is not a contact of the end-user risks preventing Telegram from identifying and preventing bad actors from continuing to perpetrate harm on the platform even after they have been blocked by an end-user on the service.

Telegram stated that the single reporting option 'Block + Report Spam' for private and Secret Chats was intended to simplify the user experience and minimise the length of time and number of interactions necessary for a user to end the chat. Telegram stated that 'once the report is processed by moderators, it is escalated as necessary – including via AI / ML if appropriate'.

eSafety notes that in response to other questions in the Notice, Telegram stated that it had no means of accessing messages reported by end-users from Secret Chats (see Section 2). Instead, Telegram stated it relies on alternative signals to assess and prioritise reports made about material in E2EE parts of the service.

eSafety notes that this may limit Telegram's ability to review, assess, prioritise, and respond to reports about harmful and illegal material or activity occurring in Telegram's Secret Chats.

**Telegram's original response to the Notice stated that end-users could make in-service reports about voice calls and video calls using a 'Violence (via the community info section)' reporting category. In response to a follow-up question from eSafety, Telegram subsequently stated that in-service reporting for voice and video calls was not available during the report period. Instead, Telegram stated that 'calls are reported together with their respective community (via the community info section and by additionally including a subset of objectionable sample messages)'.

B. Reporting of TVE by third party services that use Telegram's API

eSafety asked whether Telegram had minimum safety requirements for third party services that use Telegram's APIs to access its service. Telegram responded that it did have minimum safety requirements and that this includes the requirement for user reporting functions on third party apps to notify Telegram of breaches of its terms of service.

Telegram provided links to its Telegram API Terms of Service[196] and Security Guidelines[197], and stated that the API Terms of Service expect that third-party services make available all basic functionalities of the Telegram service, including the reporting tools, and that third-party services are accountable for ensuring that these features function correctly.

Telegram stated that it would 'from time to time re-confirm that it correctly receives user reports from the most popular third-party clients, including reports of potentially terrorist and violent content'. Telegram also stated that third-party services that fail to comply with Telegram's Terms of Service are 'routinely flagged for removal to third-party app stores and blocked from accessing Telegram's core APIs'.

[196] Telegram, 'Telegram API Terms of Service', URL supplied by Telegram on 13 September 2024, URL: https://core.telegram.org/api/terms
[197] Telegram, 'Security Guidelines', URL supplied by Telegram on 13 September 2024, URL: https://core.telegram.org/mtproto/security_guidelines

C. Reporting mechanisms for other entities to report TVE

In answer to questions about having separate reporting mechanisms for certain entities to report TVE, Telegram stated that it did have dedicated reporting mechanisms for:

- Law enforcement;
- Trusted Flaggers;
- Regulatory and public authorities; and
- 'International organizations'.

Telegram stated that these reporting mechanisms enable '[f]aster processing times whenever possible; processing by a dedicated team member / task group; deeper review if needed'. Telegram pointed to its collaboration with the Global Center for Combating Extremist Ideology, or 'Etidal', as an example of a specialised reporting mechanism. Telegram stated that between February 2022 and June 2024, Telegram removed '93,99 million pieces of TVE content' through its collaboration with Etidal.

Telegram stated that during the report period, it did not have a separate reporting mechanism for civil society groups to report TVE to the service. Telegram reported

> While engagement with these groups regarding content reporting has been minimal as at 29 February 2024, Telegram recognizes the potential benefits of collaborating with more external experts. To this end, Telegram is actively considering the introduction of dedicated contact points and other initiatives to enhance its responsiveness to valid concerns raised by civil society groups and remains open to dialogue in furtherance of that goal.

D. Percentage of TVE sent for human review

Telegram was asked to provide the percentage of TVE reports it sent for human review and the criteria and thresholds used to determine when reports were sent for human review.

Table C

Percentage of user reports of TVE sent for human review	**75%**
Criteria and thresholds used to determine when a user report is sent for human review	Telegram stated some reports were not reviewed by humans because: • Reported content had already been removed by proactive measures. • Reported content had already been removed because the channel/group it was posted in was removed.
Percentage of TVE detected through automated tools sent for human review	**65%**
Criteria and thresholds used to determine when a report of TVE is detected through automated tools is sent for human review	Telegram provided a select number of scenarios where content would be sent for human review. eSafety has chosen not publish these specific scenarios to prevent this information being misused. Telegram stated that some of these criteria were intended to prevent Telegram's systems from automatically banning 'researchers, human rights activists, legitimate news sources etc' as well to prevent bad actors from attempting to 'silence' Telegram communities by deliberately posting violative material in them as 'abusive spam'. Telegram also stated that when its automated tools detect material that is a '100% hash match', in a private Community, human moderators are notified 'but do not receive a copy of the media item itself'. In response to other questions in the Notice, Telegram stated that detections of hash matched TVE material result in resulted in the automated removal 'of all users, Communities and publications involved' except for 'Communities or users that are likely to yield false positives' (see section 5Av).

E. Percentage of TVE detected proactively

Telegram was asked what percentage of TVE was detected proactively, compared to TVE reported by users, trusted flaggers or through other channels for the following parts of its service:

Table D

Parts of Telegram	Percentage of TVE detected proactively	Percentage of TVE reported by users, trusted flaggers or other
Chats	N/A	100%
Secret Chats (E2EE)	N/A	100%
Group chats (public)	67%	33%
Group chats (private)	82%	18%
Channels (public)	69%	31%
Channels (private)	79%	21%
Voice and video calls (public and private)	N/A*	N/A*
Group video calls (public and private)	'Included in group chats'**	
Stories	60%	40%

*** In answer to a follow-up question from eSafety to clarify why its answer was 'N/A' for voice and video calls Telegram stated that voice and video calls could not be directly reported by end-users using in-service reporting tools. Instead, 'calls are reported together with their respective community (via the community info section and by additionally including a subset of objectionable sample messages)'.**

****Telegram stated that its video group call data was included in the relevant group chat statistics because 'information on resulting bans is not stored separately'. In response to other questions in the Notice, Telegram stated that it did not use any proactive detection tools to detect livestreamed TVE in group video calls. eSafety therefore understands that 100% of any TVE detected in group video calls during the report period was reported by users.**

F. Appeals against TVE-related moderation

In response to a question about how many appeals were made by users for accounts banned or content removed for TVE, where Telegram was alerted by automated tools or user reports, and how many of those were successful, Telegram provided the following information:

Table E

How Telegram was alerted to TVE	Number of appeals made for accounts banned for TVE breach	Number of appeals that were successful for accounts banned	Number of appeals made for material removed for TVE breach	Number of appeals that were successful for material removed
Automated tools	3,420	110	N/A*	
User reports	1,107	26		

***Telegram stated that because TVE-related content violations result in the users and Communities involved being removed from Telegram, 'It is not generally possible to appeal for reinstatement of removed TVE materials, so only account appeals are included'. Telegram stated that in some**

jurisdictions, such as the EU and the EU Terrorist Content Online Regulation, it may receive appeals against content removals from legally mandated contact lines. However, Telegram reported that there were 'no actionable appeals connected to removal of terrorist, violent or extremist content' during the report period.

5. Questions about proactive detection

In response to questions about the names of tools used to proactively detect known and new TVE, Telegram did not provide the names of tools used, including after eSafety asked follow-up questions seeking this information, stating that it uses an array of internal proprietary 'technical instruments' that it does not consider to be 'tools'. Telegram provided descriptions of these 'technical instruments', which it advised do not have specific names, with further information available at section 12.

In response to various questions in the Notice, Telegram stated that when TVE material was confirmed, the material was removed along with 'users, Communities and publications involved'.

A. Detecting known material using hash-matching

i. Known TVE images

In response to questions about hash matching for known TVE images, Telegram provided the following information:

Table F

Parts of service	Used image hash matching tools?	Names of tools used
Chats	No	
Secret chats (user reports)	No	
Group chats (public)	Yes	Internal Telegram Hash Matching System
Group chats (private)	Yes	
Channels (public)	Yes	
Channels (private)	Yes	
Stories	Yes	
User profile picture	Yes	
Group profile picture	Yes	
Channel profile picture	Yes	
Content in user reports	Yes	

In response to why hash matching tools were not used on Chats or Secret Chats user reports, Telegram stated that Telegram was 'founded on the principle of defending user privacy and their right to private communication' and that 'this commitment prioritizes user privacy above all'. Telegram stated that because of this commitment to user privacy

> encrypted contents of private chats are always protected, ensuring that the confidentiality of private correspondence is never compromised.

eSafety notes that Telegram stated that it does use hash-matching tools on other 'private' parts of the service – namely, private groups and private channels. eSafety further understands that Chats, Private Groups, and Private Channels all use the same form of encryption – which is not E2EE.

It is unclear to eSafety why tools capable of detecting known TVE, verified as harmful and/or violative by Telegram's own trust and safety staff, are not being used on Chats given Telegram stated that they are used on other private parts of Telegram's service, namely private groups and private channels. In relation to Secret Chats user reports, as noted at section 2B, alternative methods also exist which could enable hash-matching tools to review content reported in E2EE messages.

In response to a question about the alternative steps Telegram took to detect known TVE images on Chats and Secret Chats user reports, Telegram stated that it relied on users reporting messages via its 'Block + Report Spam' reporting tool (which as eSafety notes above, appears to only exist for messages from users that have not been added as contacts by the reporting users). Telegram stated these reports are 'processed by Telegram's tools and moderators…including via AI / ML if appropriate'. Telegram also stated that it

> employs extensive automated rate-limiting and spam-preventive measures, ensuring that no user or software is able to share content in bulk or to a significant number of users via any chat, irrespective of the nature of said content. In so doing, by design and without compromising user privacy, Telegram prevents malicious actors from effectively using its private messaging component to spread their messages.

ii. Known TVE video

In response to questions about hash matching for known TVE video, Telegram provided the following information:

Table G

Parts of service	Used image hash matching tools?	Names of tools used
Chats	No	
Secret chats (user reports)	No	
Group chats (public)	Yes	Internal Telegram Hash Matching System*
Group chats (private)	Yes	
Channels (public)	Yes	
Channels (private)	Yes	
Stories	Yes	
Content in user reports	Yes	

***In response to a follow-up question, Telegram subsequently stated that it considered that 'there is no material difference in the way in which video and image hashing is performed on a technical level'.**

In response to why hash matching tools are not used to detect known TVE videos in Chats and Secret Chats user reports, Telegram referred to its reasons for not using such tools to detect known TVE images (see section 5Ai).

In response to a question about the alternative steps Telegram took to detect known TVE videos on Chats and user reports about Secret Chats, Telegram referred to the alternative measures it took for known TVE images (see section 5Ai).

iii. Known TVE written material

In response to questions about hash matching for known TVE written material on Telegram, such as manifestos or text promoting, inciting, or instructing in TVE, Telegram provided the following information:

Table H

Parts of service	Used image hash matching tools?	Names of tools used
Chats	No	
Secret chats (user reports)	No	
Group chats (public)	No	
Group chats (private)	No	
Channels (public)	No	

Channels (private)	No	
Stories	No	
Content in user reports	Yes	Internal Telegram Hash Matching System*

***In response to a follow-up question, Telegram stated that it considered that 'there is no material difference in the way in which text and image hashing is performed on a technical level'.**

In response to why hash matching tools are only used to detect TVE written material in content referred to Telegram in user reports, Telegram stated

> Relying on hash matching on all text content at scale is generally not advisable as any one message can be expressed in numerous ways across different languages and formats. Instead, Telegram relies on ML models finetuned on known TVE written material.... Conversely, hash matching on reported content provides a useful and efficient preliminary layer to the existing moderation pipeline.

In response to a question about the alternative steps Telegram took to detect known TVE written material on chats, secret chats, private group chats, and private group channels, Telegram referred to the measures it took for known TVE images on Chats and Secret Chats (section 5Ai). For 'Other chats', which eSafety assumes refers to Public Group Chats and Public Channels, Telegram stated that it used machine learning models 'finetuned on known TVE written material to check a subset of text messages sampled from all relevant chats according to reasonable criteria'. eSafety has chosen not to publish these criteria to prevent the information being misused.

iv. Sources of TVE hashes

Telegram reported that it sourced its hashes of known TVE images, videos, and text from internal databases of hashes of TVE material that had previously been identified on Telegram and removed by its human moderators. In answer to a question about how often it updated this database, Telegram stated that the database was updated every time a human moderator removed an item of new, or previously 'unknown', TVE material from the service.

Telegram noted that its '[e]xclusive reliance' on human moderators to compile its TVE hash database is designed to mitigate risks of 'circular data pollution' posed by automated systems being trained on decisions by previous automated systems. Telegram stated that it 'purposefully avoids reintroduction of machine-labeled or synthetic data into datasets, because such measures can degrade the quality and reliability of the datasets'.

eSafety notes that limiting hash matching exclusively to material that Telegram itself has previously seen and removed risks missing TVE material that Telegram has not detected yet, and this material continuing to circulate on the platform even when such material

has already been identified by other online service providers and hashed in extensive shared databases like those run by the GIFCT or Tech Against Terrorism.

v. Action taken on known TVE

In response to questions about what action was taken when known TVE images and video were detected by its tools, Telegram stated that it resulted in the automated removal 'of all users, Communities and publications involved' except for 'Communities or users that are likely to yield false positives'. Telegram provided further information on the measures it took to address known TVE which eSafety has chosen not to publish to prevent the information being misused.

Telegram stated that when hashes of known TVE written material were detected in user reports, Telegram removed the material and the user who posted it, 'unless there is reason to believe the match may be a false positive … in which case a human moderator reviews the match and takes action accordingly.'

B. Detecting new TVE material

i. New or 'unknown' TVE images

In response to questions about the detection of new (or 'previously unknown') TVE images, Telegram provided the following information:

Table I

Parts of service	Used tools for images?	Names of tools used
Chats	No	
Secret chats (user reports)	No	
Group chats (public)	Yes	Internal Telegram AI and Machine Learning Models[198]
Group chats (private)	No	
Channels (public)	Yes	Internal Telegram AI and Machine Learning Models
Channels (private)	No	
Stories	Yes	
User profile picture	Yes	

[198] In response to a follow-up question seeking the names of the tools Telegram used to detect new forms of TVE and CSEA material, Telegram referred to the 'state-of-the-art AI/ML models that span a wide array of technologies' which it described in its original response to the Notice. These models are described at Section 12 of the Summary.

Group profile picture	Yes	Internal Telegram AI and Machine Learning Models
Channel profile picture	Yes	
Content in user reports	Yes	

When asked why it did not use technology to detect new TVE images on Chats, Secret Chats, Private Group Chats, and Private Channels, Telegram referred to the reasons it gave for not using hash matching tools to detect known TVE images (see section 5Ai).

In a follow-up question to Telegram, eSafety noted that Telegram had stated that it used tools to detect *known* TVE on Private Group Chats and Private Channels. In light of this, eSafety asked Telegram to provide the reason it did not use tools to detect *new* TVE images on these parts of the service. Telegram stated that the 'technical architecture and access rules of private groups and channels' prevent anyone who is not a member of those groups from accessing them and seeing the content being shared inside. Telegram reported that moderators could only access such content when it was reported by an end-user, or the community became accessible via a public invite link. Telegram stated that it used hash matching tools to detect known TVE on these parts of the service because when its tools detected the material, moderators would receive a notification that a 100% match with violative content had occurred – rather than the actual image or video being disclosed to them for specific review. Telegram stated that it considered that this process should not be applied to detections of new TVE material because

> Telegram contends that notifying moderators of such matches would be insufficient, as matches for new content, while accurate, cannot be verified with absolute certainty without checking the content itself. Even assuming that such matches were taken at face value, moderators would then be unable to process potential appeals by the deleted accounts.

eSafety considers that not using proactive detection tools to identify and review potential TVE material increases the likelihood that such material will remain undetected and continue to circulate on these parts of the service.

eSafety understands that chats, private group chats, and private channels are not E2EE – This means technical options are available for content detection and review by human moderators. Telegram has stated it uses such tools on other parts of its service.

In response to a question about the alternative steps Telegram took to detect new TVE images on these parts of the service, Telegram referred to the measures it took to detect likely TVE in text (see section 5Biii).

ii. New or 'previously unknown' TVE videos

In response to questions about the detection of new (or 'previously unknown') TVE videos, Telegram provided the following information:

Table J

Parts of service	Used tools for videos?	Names of tools used
Chats	No	
Secret chats (user reports)	No	
Group chats (public)	Yes	Internal Telegram AI and Machine Learning Models[199]
Group chats (private)	No	
Channels (public)	Yes	Internal Telegram AI and Machine Learning Models
Channels (private)	No	
Stories	Yes	Internal Telegram AI and Machine Learning Models
Content in user reports	Yes	

In response to why it did not use technology to detect new TVE videos on Chats, Secret Chats, Private Group Chats, and Private Channels, Telegram referred to the reasons it gave for not using proactive detection tools to detect likely TVE in images (see section 5Bi).

In response to a question about the alternative steps Telegram took to detect new TVE videos on these parts of the service, Telegram referred to measures it took to detect likely TVE in text (see section 5Biii).

iii. Text analysis to detect TVE

In response to questions about technology to detect phrases, codes or hashtags, indicating likely TVE in text (for example manifestos or text promoting, inciting, instructing TVE), Telegram provided the following information:

[199] In response to a follow-up question seeking the names of the tools Telegram uses to detect new forms of TVE and CSEA material, Telegram referred to the 'state-of-the-art AI/ML models that span a wide array of technologies' which it described in its original response to the Notice. These models are described at Section 12 of the Summary.

Table K

Parts of service	Used text analysis tools?	Names of tools used
Chats	No	
Secret chats (user reports)	No	
Group chats (public)	Yes	Internal Telegram AI and Machine Learning Models[200]
Group chats (private)	No	
Channels (public)	Yes	Internal Telegram AI and Machine Learning Models
Channels (private)	No	
Stories	Yes	Internal Telegram AI and Machine Learning Models
Profile username	Yes	
Profile description	Yes	
Group username	Yes	
Group description	Yes	
Channel username	Yes	
Channel description	Yes	
Content in user reports	Yes	

In response to why it did not use technology to scan Chats, Secret Chats, Private Group Chats, and Private Channels for indications of likely TVE in text, Telegram referred to the reasons it gave for not using hash matching tools to detect known TVE images (see section 5Ai).

In response to what alternative steps Telegram took to detect known phrases, codes, or hashtags indicating likely TVE on these parts of the service, Telegram stated that it provided reporting options for users to report such material on these parts of the service. Telegram noted that when users report such TVE material in 'private groups and channels', this results in the material being forwarded to Telegram moderators for their review. However, for what Telegram describes as 'private 1-on-1 chats' (i.e. chats and Secret Chats), Telegram reiterated that user reports in these parts of the service are 'processed by Telegram's tools and moderators…including via AI / ML if appropriate'.

Telegram also referred to the 'automated rate limiting and spam-preventative measures' it used to prevent bad actors from spreading known TVE material in private messages (see section 5Ai).

[200] In response to a follow-up question seeking the names of the tools Telegram used to detect new forms of TVE and CSEA material, Telegram referred to the 'state-of-the-art AI/ML models that span a wide array of technologies' which it described in its original response to the Notice. These models are described at Section 12 of the Summary.

iv. Sources of phrases, codes, and hashtags

Telegram stated that it sourced phrases, codes, and hashtags indicating likely TVE from 'samples … based on content items' that had been removed from Telegram by its human moderators.

v. Action taken when new TVE was detected

In response to questions about what action was taken when new TVE images, video, and likely TVE in text was detected by its tools, Telegram stated that the material was sent for human review and if the content was confirmed as TVE it resulted in the removal of 'users, Communities and publications involved'. Telegram reported that following removal, new TVE images and video were then added to Telegram's internal hash database so they could be actioned in the future as known TVE.

Telegram provided further information on the measures it took to address new TVE which eSafety has chosen not to publish to prevent the information being misused.

Telegram also noted that when a Community is removed, human moderators review the most common search terms that were used to find that Community in order for them to be 'possibly removed from Telegram's public search to limit future spread and reach of similar content'.

In answer to a question asking if the detection of phrases, codes, or hashtags indicating likely TVE in text resulted in Telegram blocking these words or phrases to users searching for them, Telegram responded that 'yes', it blocked 'certain keywords or text patterns' from its search results.

C. Languages covered by language analysis tools

i. Detecting TVE in text

In response to questions about the languages covered by Telegram's language analysis tools, Telegram did not provide a list of languages.

Telegram stated that its models and tools for detecting TVE text 'perform reasonably well in most languages', and that these technologies 'aim to abstract away the concept of language when deriving embeddings from text'.

In response to follow-up questions from eSafety, Telegram stated

> Telegram did not maintain a specific list of all languages included in the training sets of the underlying models, which is why it was unable to provide a list relevant for the report period. Telegram must note that a significant share of messages in its datasets contain text in

multiple languages (i.e., multiple languages within the same message). As well, messages are often too short to automatically identify a specific language with absolute certainty.

ii. Detecting TVE in video

In response to a question about the languages covered by the tools Telegram used to detect new TVE in video, Telegram did not provide a list of languages.

Telegram stated the tools did not 'have particular regard for the specific language of the content in question' but focussed on detecting patterns the model had previously learned to associate with TVE from an existing dataset.

In response to follow-up questions from eSafety, Telegram referred to the answer it gave to eSafety's follow-up question regarding the languages covered by the tools Telegram used to detect likely TVE in text (see section 5Ci).

D. Livestreamed TVE

i. Detecting livestreamed TVE

The Notice specified that livestreaming includes one-on-one video calls and video calls where one or more multiple people stream material to a group of any size.

In response to questions about the measures Telegram had in place to detect the livestreaming of TVE on its service, Telegram provided the following information:

Table L

Parts of service	Measures in place to detect TVE in livestreams?	Interventions used	Name of tools used
Group video calls	No		
Channel livestreams	No		

When asked why it did not have any measures in place to detect livestreamed TVE, Telegram stated

> While livestreaming functionalities are supported on Telegram, they represent a generally insignificant share of the service's overall usage, particularly so as it concerns the spread of harmful content. As such, Telegram finds that immediate user reports already provide reliable coverage to detect and address such incidents effectively.

ii. Reducing the likelihood of livestreamed TVE

In response to questions about the steps taken by Telegram to reduce the likelihood that TVE could occur in livestreams, Telegram stated that it used the following measures:

- Restrictions for those who have previously violated terms of service or community guidelines/standards.
- User reports – Telegram stated that it 'relies on immediate user reports' to detect livestreamed TVE **but did not indicate whether or how it prioritises reviews of reports of livestreamed content.** In response to other questions in the Notice, Telegram stated that it did not provide in-service reporting tools for video calls (see section 4A) during the report period.

Telegram stated that livestreaming functionalities represent an 'insignificant share of the service's overall usage', particularly so as it concerns the spread of harmful content. As such, Telegram finds that immediate user reports already provide reliable coverage to detect and address such incidents effectively.'

E. Blocking links to TVE material

i. Detection and sources of URLs

Telegram was asked about its use of lists or databases to proactively detect and block URLs linking to TVE on other platforms. Specifically, Telegram was asked about:

- Known URLs linking to websites/services operated by individuals/organisations dedicated to the creation, promotion, or dissemination of TVE
- URLs linking to known TVE material on other services/websites (which may not be dedicated to TVE)
- Join-links to groups, Channels, communities, or forums on other services that were known to be associated with TVE.

Table M

Parts of service	Blocked URLs to websites/services dedicated to TVE?	Blocked URLs linking to known TVE material on other services/websites?	Blocked join-links to groups/channels on other services known to be associated with TVE?	URL sources
Chats	No	No	No	
Secret chats (E2EE)	No	No	No	
Group chats (public)	No	No	No	
Group chats (private)	No	No	No	
Channels (public)	No	No	No	
Channels (private)	No	No	No	
Profile description	No	No	No	
Group description	No	No	No	
Channel description	No	No	No	

When asked why URLs to TVE material were not blocked and whether alternative steps were taken to block URLs, Telegram stated that 'focusing its efforts on ML-based classification tends to yield better results when compared to static link blacklists'. Telegram stated that links to harmful material tended to be 'routinely taken down by all hosting providers and tend to either rotate constantly or be hidden behind URL shorteners, proxies etc'. Telegram also stated that it used Internal Telegram AI and Machine Learning Models (see Section 12) that are trained on previous detections of TVE material, including material that may contain external links.

F. Off-platform monitoring

Telegram was asked if it used off-platform monitoring[201] either provided internally or by third-party services, to identify accounts, groups or channels on its service that were dedicated to TVE. Telegram stated it performs '[e]xtensive monitoring' of media sources as well as reviewing referrals sent by 'non-registered users and trusted organizations' to Telegram via email.

[201] Monitoring of activity on other services.

6. Questions about resources, expertise, and human moderation

A. Trust and Safety

i. Trust and Safety and other staff

Telegram was asked to provide the number of staff that were employed or contracted by Telegram to carry out certain functions at the end of the report period.

Table N

Category of staff	Number of staff*
Engineers employed by Telegram focussed on trust and safety	5
Content moderators employed by Telegram	0
Content moderators contracted by Telegram	150
Trust and safety staff employed by Telegram (other than engineers and content moderators)	4

*** Telegram stated that these figures represented the number of staff who 'may from time to time be involved with decisions regarding content or reports from Australia and do not reflect or approximate the total number of global content moderation and trust and safety personnel contracted by Telegram.' Telegram also stated that Australian end-users make up less than 0.2% of its monthly active users.**

eSafety notes that Telegram did not provide its global resourcing. In response to follow-up questions from eSafety seeking the total numbers of staff in the categories, rather than the numbers that 'may from time to time be involved with decisions regarding content or reports from Australia', Telegram did not provide the information.

eSafety notes that TVE is a global harm and the resources that a service has in place to respond to 'content or reports' internationally is highly relevant to the online safety of Australians, and implementation of the Expectations.

ii. Trust and Safety dedicated to minimising TVE

In response to a question asking if Telegram had a dedicated trust and safety team responsible for minimising TVE on the service, Telegram answered 'yes', reporting that '[t]he relevant team members are high-performing professionals trained in team management, threat mitigation and legal affairs'. Telegram stated that this team was responsible for 'overseeing content moderation and reviewing reports from external stakeholders, such as trusted flaggers and international organizations'.

Telegram provided the following information about the composition of its team:

Table O

Name of role/area of expertise	Number of staff	Number of contractors
Trust and Safety managers*	4**	0

*** Telegram stated that it did not have specific titles for these positions because Telegram's 'hierarchy is informal and horizontal'.**

**** Telegram stated that this figure was specific to 'staff that may from time to time be involved in decisions regarding content or reports from Australia and do not reflect or approximate the total number of global trust and safety personnel contracted by Telegram'.**

In response to follow-up questions from eSafety seeking the total numbers of staff in the categories, rather than the numbers that 'may from time to time be involved with decisions regarding content or reports from Australia', Telegram did not provide the information.

iii. Surge teams to respond to a TVE crisis

Telegram was asked if it had a surge team(s) to respond to TVE crises, such as a livestreamed attack with content disseminated on the service. Telegram answered 'yes' and stated that it 'maintains crisis mitigation protocols in accordance with industry standards'. Telegram stated that on occasion it could establish task forces of '2-3 people tailored to resolving specific situations, e.g., appearance of terrorist or violent content in significant quantities'.

Telegram provided the following information about the composition of this team:

Table Q

Name of role/area of expertise	Number of staff	Number of contractors
Trust and Safety managers and contractors*	3**	13**

*** Telegram stated that it did not have specific titles for these positions because Telegram's 'hierarchy is informal and horizontal'.**

**** Telegram stated that these figures were specific to 'staff that may from time to time be involved in decisions regarding content or reports from Australia and do not reflect or approximate the total number of global trust and safety personnel contracted by Telegram'.**

In response to follow-up questions from eSafety seeking the total numbers of staff in the categories, rather than the numbers that 'may from time to time be involved with decisions regarding content or reports from Australia', Telegram did not provide the information.

B. Languages human moderators operate across

In response to a question about the languages that its human moderators operated across (both employees and contractors), Telegram provided the following:

Table R

Languages covered by employees (all languages)	Languages covered by contractors (all languages)[202]	
N/A*	• English • Amharic • Arabic • Azerbaijani • Bulgarian • Chinese (traditional and simplified) • Croatian • Czech • Danish • Estonian • Farsi • Filipino • Finnish • French • Georgian • German • Greek • Hindi • Icelandic • Indonesian • Italian • Japanese	• Kazakh • Korean • Kyrgyz • Luganda • Lunyakore • Lusoga • Malay • Moldavian • Norwegian • Polish • Portuguese (Brazil) • Portuguese (Europe) • Romanian • Russian • Serbian • Shona • Spanish • Swahili • Swedish • Tajik • Turkish • Ukrainian • Urdu • Uzbek • Yoruba

[202] Telegram also advised that since the report period, it had expanded the languages covered by its contracted content moderators by adding Afrikaans, Bengali (Bangladesh), Chichewa (Zambia), Dhivehi (Maldives), Dutch, Gujarati, Kabyle (Algeria), Kinyarwanda, Lithuanian, Macedonian, Punjabi, Sinhalese (Sri Lanka), and Thai.

***Telegram stated that '[a]ll ordinary moderators' on Telegram are contractors.**

eSafety notes that the top 5 languages, other than English, spoken in Australian homes are Arabic, Cantonese, Mandarin, Vietnamese and Punjabi.[203] Telegram's human moderators do not cover Vietnamese or Punjabi.[204]

C. Median time to reach an outcome to a user report of TVE

Telegram was asked to provide the median time taken to reach an outcome after receiving a user report about TVE for the following parts of the service:

Table S

Parts of the service	Reports from users globally	Reports from users in Australia *
Chats	18 hours	18 hours
Secret Chats	18 hours	18 hours
Group chats (public)	15 hours	15 hours
Group chats (private)	15 hours	15 hours
Channels (public)	15 hours	15 hours
Channels (private)	15 hours	15 hours

In response to a question asking how median time was calculated Telegram stated that to calculate these figures it registered 'the net time frames between the submission of each individual report and the moderator's decision in respect of that report'.

*** Telegram stated that it 'currently doesn't have the technical means to provide separate statistics by country'.**

It is unclear to eSafety how Telegram determines that it has reached an outcome for Secret Chats when it has stated that it does not review the contents of the messages being reported on this part of the service (see section 2B).

C. Volunteer moderation

In response to questions about the process its volunteer moderators followed, and the processes Telegram had in place to monitor their conduct and uphold moderation standards

[203] Australian Bureau of Statistics, 'Cultural diversity: Census', 28 June 2021, URL: https://www.abs.gov.au/statistics/people/people-and-communities/cultural-diversity-census/latest-release#:~:text=Top%205%20languages%20used%20at,Punjabi%20(0.9%20per%20cent).

[204] Following consultation with Telegram on the proposed report for publication, Telegram noted that Punjabi was added to its list of covered languages since the report period as per footnote 203.

Telegram responded that it relied on 'contracted professional moderators' and did not have volunteer moderators as at 29 February 2024. Following consultation with Telegram on the proposed report for publication, Telegram noted that it had interpreted eSafety's definition of 'volunteer moderator' differently and updated its response to provide the following information:

Table T

Question	Details provided by Telegram
Did Telegram have a standards policy, or similar, outlining the responsibilities and expectations of volunteer moderator roles?	Yes Telegram reported that a 'collection of short instructions and overviews related to group management and moderation features' 'can…from time to time' be hosted on the Telegram website or provided as in-app guidance to group administrators.[205] Telegram noted that only group administrators operate as 'volunteer moderators' as they are able to moderate content published by other group members whereas in channels only channel administrators can publish content, not channel subscribers, therefore channel administrators are not considered to be volunteer moderators. Telegram also noted that group administrators moderate user comments on channel publications as user comments 'are technically made in associated *groups*'. Telegram noted that '[i]f the group administrators do not directly violate Telegram's Terms of Service (e.g., when a group was temporarily abused by malicious users), Telegram's moderators may at their own discretion temporarily close it allowing the group administrators the opportunity to address any violations.' However, Telegram reported that if a community is being used to share 'illicit content' whether by Community administrators or community members, the community, administrators and users who are in violation of Telegram's terms of service may be permanently terminated.

[205] Telegram.org, 'Supergroups 10,000: Admin Tools & More' https://telegram.org/blog/admin-revolution. 'Aggressive Anti-Spam' https://telegram.org/blog/ultimate-privacy-topics-2-0#aggressive-anti-spam.
'Groups – Admin Tools' https://telegram.org/tour/groups#admin-tools
'Join Requests for Groups and Channels' https://telegram.org/blog/shared-media-scrolling-calendar-join-requests-and-more
'Slow Mode' https://telegram.org/blog/silent-messages-slow-mode#slow-mode
'Group Permissions' https://telegram.org/blog/permissions-groups-undo

What training and/or guidance was provided to volunteer moderators regarding proactive minimisation of TVE and removal of accounts that share TVE.	Telegram reported that 'the primary focus of the group administrators lies in addressing abusive spam in their groups' but that the information and URLs provided in the above response cover all harm types, not just TVE.
Were users able to make in-service reports about volunteer moderators in instances where they were failing to meet any required responsibilities and expectations?	Telegram responded 'Yes' Telegram's response indicated that a user can report the Community in-service. It did not indicate that a specific report about a volunteer moderator can be made in-service.
If volunteer moderators removed an account from a public channel, private channel, or a group for TVE-breaches, were trust and safety staff informed?	Telegram responded 'Yes' that trust and safety staff are informed when a volunteer moderator removes an account from a public channel, private channel or group for TVE breaches. However, Telegram's response stated that its administrators '**may**' opt to report the removal of 'a user or their messages (in whole or in part) from a group' to Telegram with a detailed description of the infringement. eSafety understands that Telegram trust and safety are therefore not automatically informed when a volunteer moderator removes an account. Telegram stated that its 'systems can track when, why and how often a user was removed by group administrators and may escalate matters accordingly on a case-by-case basis. Additionally, these indicators are considered by Telegram's AI models, both to prioritize reports and to take action autonomously.'
If Telegram's Trust and Safety staff banned a user for a TVE-related violation in a Community, were the volunteer moderators of that group or channel notified?	Telegram responded 'Yes' that volunteer moderators are informed when Telegram's trust and safety staff banned a user for a TVE-related violation in a Community. Telegram stated that its 'systems are programmed to notify group administrators in cases where the user's publications were removed from the group, even if the user itself was not banned.'

'Mass Moderation for Groups' (made available after the report period) https://telegram.org/blog/my-profile-and-15-more/ru?setln=en#mass-moderationfor-groups.
T.me 'Telegram Tips' https://t.me/s/TelegramTips/115, https://t.me/s/TelegramTips/333, https://t.me/s/TelegramTips/380, https://t.me/s/TelegramTips/447.
URLs provided by Telegram on 21 December 2024.

7. Questions about steps to prevent recidivism

A. Measures and indicators

In response to a question asking if Telegram had measures in place to prevent recidivism for TVE-related breaches on its service, Telegram responded 'yes' and **listed a minimal[206] number of indicators that it used to detect users that have previously been banned for TVE breaches.** eSafety has chosen not to publish these indicators to prevent the information being misused.

B. Preventing banned TVE groups and channels from being recreated

In response to a question about the measures Telegram took to prevent banned TVE groups and channels from being recreated, Telegram stated that '[o]wners and administrators of infringing Communities may also face removal alongside the Communities themselves, preventing them from creating new Communities or accounts on Telegram'. Telegram also referred to a minimal number of signals it may use to 'routinely detect' Communities that bear similarities to previously banned Communities which eSafety has chosen not to publish to prevent the information being misused.

C. Applying TVE-related bans to associated accounts

Telegram was asked, when it took action against an account for a TVE-related breach, whether it applied bans to associated accounts. eSafety defined 'associated accounts' as 'other users who are associated with the banned user'. Telegram answered 'yes' and stated that when it identified a user disseminating TVE material, it reviewed 'further reports linked to this user, as well as to any Communities which the user owns or administrates'. Telegram stated that any Communities found to be involved in disseminating TVE would also be removed.

Telegram stated that Channel subscribers or group members 'who are neither managing nor directly spreading or promoting prohibited content, even if they are part of Communities that may contain such content, are not subject to automatic bans'. Telegram stated that this approach was adopted to avoid inadvertently disrupting law enforcement, journalists, activists and others who may be part of these groups for legitimate reasons.

Telegram further reported that while it 'does not compile data or operate tools conducive to internal investigative work on private user interactions', Telegram has developed tools to assist

---

[206] eSafety uses the following terms to give an impression of the extent of the indicators used by services in the table below, rather than publishing the specific indicators which could be misused:
- Minimal: A small number
- Several: A moderate number
- Multiple: A significant number.

its moderators to identify 'interconnected management' between 'flagged Communities and their administrators'. eSafety has chosen not to publish further detail on the measures Telegram reported to prevent the information being misused.

D. Sharing of banned account details with other entities

Telegram was asked if it shared details of accounts banned for TVE with the following entities:

Table U

Entity	Shared details of accounts banned for TVE?
Other service providers	No
Law enforcement	Yes*
Regulatory or other public authorities	No
Global Internet Forum to Counter Terrorism	No
Civil society groups	No

*** Telegram stated that it provided information to law enforcement in response to 'valid legal requests submitted by law enforcement agencies through designated channels'.**

Telegram stated that as at 29 February 2024, Telegram had not received any legal requests from Australian law enforcement agencies 'via dedicated formal channels (e.g., mutual legal assistance requests to the governments in jurisdictions in which the relevant Telegram companies are located)'.

Part 2. Questions in relation to chid sexual exploitation and abuse (CSEA)

1. Questions about reporting of CSEA

A. In-service reporting of CSEA on different parts of Telegram

In response to questions about whether users could report instances of CSEA to Telegram within the service (as opposed to navigating to a separate webform or email address), Telegram responded:

Table V

Parts of the service	In-service reporting option?	Reporting category
Chats	Yes	'Block user > Report Spam'
Secret Chats	Yes	
Group chats (public)	Yes	'Child Abuse'
Group chats (private)	Yes	
Channels (public)	Yes	
Channels (private)	Yes	
Voice calls	No*	
Video calls	No*	
Stories	Yes	'Child Abuse'

*Telegram's original response to the Notice stated that end-users could make in-service reports about voice calls and video calls using a 'Child Abuse (via the Community's info section)' reporting category. In response to a follow-up question from eSafety, Telegram subsequently stated that in-service reporting for voice and video calls was not available during the report period. Instead, Telegram stated that 'calls are reported together with their respective community (via the community info section and by additionally including a subset of objectionable sample messages)'.

As noted at section 4Ai, Telegram subsequently clarified that the 'Block + Report Spam' reporting flow is only available when the Chat or Secret Chat is 'initiated by non-contacts and strangers'. eSafety understands that when an end-user wishes to report a message from an account they have already added as contact, the only option in-service is to 'Block user'.

With respect to TVE (which eSafety understands is applicable to CSEA), Telegram stated that this was because

> In the extremely unlikely event that a user's friend or acquaintance began sending them TVE content, Telegram contends that it would be more reasonable and effective for said user to contact authorities directly, providing all relevant proof and contact information.

As noted at section 4Ai, eSafety considers that limiting reporting tools to scenarios where the account sending harmful or violative material is not a contact of the end-user risks preventing Telegram from identifying and preventing bad actors from continuing to perpetrate harm on the platform even after they have been blocked by an end-user on the service.

Telegram stated that the single reporting option 'Block + Report Spam' for private and Secret Chats was intended to simplify the user experience and minimise the length of time and number of interactions necessary for a user to end the chat. Telegram stated that 'once the report is processed by moderators, it is escalated as necessary – including via AI / ML if appropriate'.

As noted at section 4A, eSafety notes that in response to other questions in the Notice, Telegram stated that it had no means of accessing messages reported by end-users from Secret Chats (see Section 2). Instead, Telegram stated it relies on alternative signals to assess and prioritise reports made about material in E2EE parts of the service.

eSafety notes that this may limit Telegram's ability to review, assess, prioritise, and respond to reports about harmful and illegal material or activity occurring in Telegram's Secret Chats.

2. Questions about proactive detection of CSEA

In response to questions about the names of tools used to proactively detect known and new CSEA, Telegram did not provide the names of tools used, including after eSafety asked follow-up questions seeking this information, stating that it uses an array of internal proprietary 'technical instruments' that it does not consider to be 'tools'. Telegram provided descriptions of these 'technical instruments', which it advised do not have specific names, with further information available at section 12.

In response to various questions in the Notice, Telegram stated that when CSEA material was confirmed, the material was removed along with 'users, Communities and publications involved'.

A. Detecting known material using hash matching

i. Known CSEA images

In response to questions about hash matching for known CSEA images, Telegram provided the following information:

Table W

Parts of service	Used image hash matching tools?	Names of tools used
Chats	No	
Secret chats (user reports)	No	
Group chats (public)	Yes	Internal Telegram Hash Matching System
Group chats (private)	Yes	
Channels (public)	Yes	
Channels (private)	Yes	
Stories	Yes	
User profile picture	Yes	
Group profile picture	Yes	
Channel profile picture	Yes	
Content in user reports	Yes	

In response to why hash matching tools were not used on Chats or Secret Chats user reports, Telegram referred to its reasons for not using such tools to detect known TVE images (see section 5Ai).

In response to a question about the alternative steps Telegram took to detect known CSEA images on Chats and Secret Chats user reports, Telegram referred to the alternative measures it took for known TVE images (see section 5Ai). Telegram also noted that it

> maintains a dedicated hotline StopCA@telegram.org for reporting any content related to CSAM.

ii. Known CSEA video

In response to questions about hash matching for known CSEA video, Telegram provided the following information:

Table X

Parts of service	Used image hash matching tools?	Names of tools used
Chats	No	
Secret chats (user reports)	No	
Group chats (public)	Yes	Internal Telegram Hash Matching System
Group chats (private)	Yes	
Channels (public)	Yes	
Channels (private)	Yes	
Stories	Yes	
Content in user reports	Yes	

When asked why hash matching tools were not used to detect known CSEA videos in Chats and user reports about Secret Chats, Telegram referred to its reasons for not using such tools to detect known TVE images (see section 5Ai).

In response to a question about the alternative steps Telegram took to detect known TVE videos on Chats and user reports about Secret Chats, Telegram referred to the alternative measures it took for known TVE images and known CSEA images (see section 5Ai).

iii. Sources of CSEA hashes

Telegram reported that it sourced its hashes of known CSEA images and videos from internal databases of hashes of CSEA material that had previously been identified on Telegram and removed by its human moderators. In answer to a question about how often it updated this database, Telegram stated that the database was updated every time a human moderator removed an item of new, or previously 'unknown', CSEA material from the service.

Telegram also referred again to the reasons it gave for its 'exclusive reliance' on human moderators to compile its TVE hash database (see section 5Aiv).

eSafety notes that limiting hash matching exclusively to material that Telegram itself has previously seen and removed risks missing CSEA material that Telegram has not detected yet, and this material continuing to circulate on the platform even when such material has already been identified by other online service providers and hashed in extensive shared databases like those run by the IWF or NCMEC.

This means that to the extent that it used hash matching tools, Telegram did not have access to NCMEC's hash database, which contains more than 5 million hashes of verified

CSEA material.[207] eSafety notes media reporting that NCMEC and the IWF both claimed to have made past efforts to contact Telegram that went ignored prior to the CEO of Telegram's arrest on 27 August 2024.[208]

iv. Action taken when CSEA hashes are detected

Telegram stated that detections of known CSEA images and videos through hash-matching resulted in the automated removal 'of all users, Communities and publications involved'. Telegram also referred to further information it had provided on the measures it took to address known TVE which eSafety has chosen not to publish to prevent the information being misused.

B. Detecting new CSEA material

i. Text analysis to detect CSEA

In response to questions about technology to detect terms, abbreviations, codes and hashtags indicating likely CSEA (for example grooming, sexual extortion, or the trading and sale of CSEA material), Telegram provided the following information:

Table Y

Parts of service	Used text analysis tools?	Names of tools used
Chats	No	
Secret chats (user reports)	No	
Group chats (public)	Yes	Internal Telegram AI and Machine Learning Models[209]
Group chats (private)	No	
Channels (public)	Yes	Internal Telegram AI and Machine Learning Models
Channels (private)	No	
Stories	Yes	
Profile username	Yes	
Profile description	Yes	
Group username	Yes	

[207] Google Safety Center, 'NCMEC, Google and Image Hashing Technology', accessed 15 November 2024, URL: https://safety.google/stories/hash-matching-to-help-ncmec/

[208] NBC News, 'Telegram ignored outreach outreach from child safety watchdogs before CEO's arrest, groups say', 28 August 2024, URL: https://www.nbcnews.com/tech/security/telegram-ceo-pavel-durov-child-safety-rcna168266

[209] In response to a follow-up question seeking the names of the tools Telegram uses to detect new forms of TVE and CSEA material, Telegram referred to the 'state-of-the-art AI/ML models that span a wide array of technologies' which it described in its original response to the Notice. These models are described at Section 12 of the Summary.

Group description	Yes	Internal Telegram AI and Machine Learning Models
Channel username	Yes	
Channel description	Yes	
Content in user reports	Yes	

In response to why it did not use technology to scan Chats, Secret Chats, private group chats, and private channels for indications of likely CSEA, Telegram referred to its reasons for not using such tools to detect known TVE images (see section 5Ai).

In response to what alternative steps Telegram took to detect known terms, abbreviations, codes or hashtags indicating likely CSEA on these parts of the service, Telegram referred to the alternative steps it took to detect text indicating likely TVE (see section 5Bi). Telegram also referred to the dedicated hotline it maintains for receiving reports about CSAM (see section 9Ai).

ii. Sources of terms, abbreviations, codes, and hashtags

Telegram stated that it sourced phrases, codes, and hashtags indicating likely CSEA from samples of CSEA material that had been removed from Telegram by its human moderators.

iii. Languages covered by language analysis tools

When asked what languages were covered by technology used to detect terms, abbreviations, codes and hashtags indicating likely CSEA, Telegram did not provide a list of languages.

Telegram referred to the answer it gave in response to questions about the languages covered by the tools Telegram used to detect likely TVE in text (see section 5Ci).

iv. New or 'unknown' CSEA images

In response to questions about the detection of new (or 'previously unknown') CSEA images, Telegram provided the following information:

Table Z

Parts of service	Used tools for images?	Names of tools used
Chats	No	
Secret chats (user reports)	No	
Group chats (public)	Yes	Internal Telegram AI and Machine Learning Models[210]
Group chats (private)	No	
Channels (public)	Yes	Internal Telegram AI and Machine Learning Models
Channels (private)	No	
Stories	Yes	Internal Telegram AI and Machine Learning Models
User profile picture	Yes	
Group profile picture	Yes	
Channel profile picture	Yes	
Content in user reports	Yes	

In response to why it did not use technology to detect new CSEA images on Chats, Secret Chats, Private Group Chats, and Private Channels, Telegram referred to the reasons it gave for not using proactive detection tools to detect new TVE images (see section 5Bi).

In response to a question about the alternative steps Telegram took to detect new CSEA images on these parts of the service, Telegram referred to measures it took to detect known terms, abbreviations, codes and hashtags that indicate likely CSEA on the service and likely TVE in text (see sections 9Bi and 5Bi).

As noted at section 5Bi, eSafety considers that not using proactive detection tools to identify and review potential CSEA material increases the likelihood that such material will remain undetected and continue to circulate on these parts of the service.

eSafety understands that Chats, Private Group chats, and Private Channels are not E2EE – leaving alternative technical options available for content detection and review by human moderators.

v. New or 'previously unknown' CSEA videos

In response to questions about the detection of new (or 'previously unknown') CSEA videos, Telegram provided the following information:

[210] In response to a follow-up question seeking the names of the tools Telegram used to detect new forms of TVE and CSEA material, Telegram referred to the 'state-of-the-art AI/ML models that span a wide array of technologies' which it described in its original response to the Notice. These models are described at Section 12 of the Summary.

Table AA

Parts of service	Used tools for videos?	Names of tools used
Chats	No	
Secret chats (user reports)	No	
Group chats (public)	Yes	Internal Telegram AI and Machine Learning Models[211]
Group chats (private)	No	
Channels (public)	Yes	Internal Telegram AI and Machine Learning Models
Channels (private)	No	
Stories	Yes	Internal Telegram AI and Machine Learning Models
Content in user reports	Yes	

In response to why it did not use technology to detect new CSEA videos on Chats, Secret Chats, Private Group Chats, and Private Channels, Telegram referred to the reasons it gave for not using proactive detection tools to detect new TVE images (see section 5Bi).

In response to a question about the alternative steps Telegram took to detect new CSEA videos on these parts of the service, Telegram referred to measures it took to detect known terms, abbreviations, codes and hashtags that indicate likely CSEA on the service and likely TVE in text (see sections 9Bi and 5Bi).

vi. Action taken when new CSEA is detected

Telegram stated that when likely new CSEA material was detected, it was either immediately processed by Telegram's automated tools or sent for human review 'depending on the degree of confidence to which the relevant model is able to issue a judgement, combined with other factors'. eSafety has chosen not to disclose these additional factors due to public safety reasons. If the detection was confirmed, it resulted in the removal of all 'users, Communities and publications involved'. Telegram reported that when new CSEA images and videos were removed, they were then added to Telegram's internal hash database.

C. Blocking links to CSEA material

i. URLs linking to known CSEA

In response to a question about whether Telegram blocked URLs linking to known CSEA, Telegram provided the following information:

[211] In response to a follow-up question seeking the names of the tools Telegram used to detect new forms of TVE and CSEA material, Telegram referred to the 'state-of-the-art AI/ML models that span a wide array of technologies' which it described in its original response to the Notice. These models are described at Section 12 of the Summary.

Table BB

Parts of service	Blocked URLs linking to known TVE material on other services/websites?	URL sources
Chats	No	
Secret chats (E2EE)	No	
Group chats (public)	No	
Group chats (private)	No	
Channels (public)	No	
Channels (private)	No	
Profile description	No	
Group description	No	
Channel description	No	

In response to why URLs to known CSEA material were not blocked and whether alternative steps were taken to block such URLs, Telegram reiterated that 'focusing its efforts on ML-based classification tends to yield better results when compared to static link blacklists'. Telegram stated that links to harmful material tend to be taken down routinely or hidden behind URL shorteners. Telegram stated that it used proactive detection tools that are trained on previous detections of CSEA material, including material that may contain external links.

Telegram also stated that, as at October 2024, it was 'in the process of joining the Internet Watch Foundation's safety programs involving *inter alia* access to URL lists containing links to known CSAM websites'.

As noted above, eSafety is aware of public statements made by the Internet Watch Foundation (IWF) asserting that prior to the arrest of Telegram's CEO in August 2024, the IWF had made repeated efforts to reach out to Telegram and that Telegram had refused to 'take any of its services to block, prevent, and disrupt the sharing of child sexual abuse imagery'.[212] It is unclear why Telegram did not take the opportunity to work with the IWF sooner.

D. Percentage of CSEA detected proactively

Telegram was asked what percentage of CSEA was detected proactively, compared to CSEA reported by users, trusted flaggers or through other channels for the following parts of its service:

[212] NBC News, 'Telegram ignored outreach outreach from child safety watchdogs before CEO's arrest, groups say', 28 August 2024, URL: https://www.nbcnews.com/tech/security/telegram-ceo-pavel-durov-child-safety-rcna168266.

Table CC

Parts of Telegram	Percentage of CSEA detected proactively	Percentage of CSEA reported by users, trusted flaggers or other
Chats	N/A	100%
Secret Chats	N/A	100%
Group chats (public)	71%	29%
Group chats (private)	85%	15%
Channels (public)	74%	26%
Channels (private)	80%	20%
Voice and video calls (public and private)	N/A*	N/A*
Group video calls (public and private)	'Included in group chats'**	
Stories	65%	35%

*** In answer to a follow-up question from eSafety to clarify why its answer was 'N/A' for voice and video calls, Telegram referred to the answer it gave with respect to the percentage of TVE detected proactively and by reports on voice and video calls (see section 4E).**

****Telegram stated that its group video call data was included in the relevant group chat statistics because 'information on resulting bans is not stored separately'.**

E. Appeals against CSEA-related moderation

In response to a question about how many appeals were made by users for accounts banned or content removed for CSEA, where Telegram was alerted by automated tools or user reports, and how many of those were successful, Telegram provided the following information:

Table DD

How Telegram was alerted to CSEA	Number of appeals made for accounts banned for CSEA breach	Number of appeals that were successful for accounts banned	Number of appeals made for material removed for CSEA breach	Number of appeals that were successful for material removed
Automated tools	7,098	573	N/A*	
User reports	2,702	218		

***Telegram stated that because CSEA-related content violations resulted in the users and Communities involved being removed from Telegram, 'It is not generally possible to appeal for reinstatement of removed CSAM materials, so only account appeals are included'. Telegram stated that in some jurisdictions, such as the EU and the EU Terrorist Content Online Regulation, it may receive appeals**

against content removals from legally mandated contact lines. However, Telegram reported that there were 'no such appeals connected to removal of CSAM content' during the report period.

3. Questions about resources, expertise, and human moderation

A. Median time to reach an outcome to a user report of CSEA

Telegram was asked to provide the median time taken to reach an outcome[213] after receiving a user report about CSEA for the following parts of the service:

Table EE

Parts of the service	Reports from users globally	Reports from users in Australia *
Chats	11 hours	11 hours
Secret Chats	11 hours	11 hours
Group chats (public)	10 hours	10 hours
Group chats (private)	10 hours	10 hours
Channels (public)	10 hours	10 hours
Channels (private)	10 hours	10 hours

In response to a question asking how median time was calculated Telegram stated that to calculate these figures it registered 'the net time frames between the submission of each individual report and the moderator's decision in respect of that report'.

*** Telegram stated that it 'currently doesn't have the technical means to provide separate statistics by country'.**

4. Questions about steps to prevent recidivism

A. Measures and indicators

In response to a question asking if Telegram had measures in place to prevent recidivism for CSEA-related breaches on its service, Telegram responded 'yes' and stated:

> Given the severity of CSAM, any infringement related to it typically results in the permanent removal of related accounts and Communities. Owners of infringing groups and channels

[213] Defined in the Notice as a calculation from 'the time that a user report is made, to a content moderation outcome or decision, such as removing the content, banning the account, or deciding that no action should be taken.'

may also face removal, preventing them from creating new Communities or accounts on Telegram.

Telegram listed a minimal[214] number of indicators that it used to detect users that have previously been banned for CSEA breaches. eSafety has chosen not to publish these indicators to prevent the information being misused.

5. Additional information

In response to an opportunity to provide further information and context to any of its responses to the questions asked in the Notice, Telegram added that:

> Telegram was built to safeguard the privacy of individuals at risk – such as legitimate activists, journalists, and protesters – and preserve their right to private correspondence. While staying true to its core value of user privacy, Telegram actively engages in policy efforts and implements robust moderation tools to address abusive content.
>
> Telegram's exponential growth in the recent years has presented unique moderation challenges due to the sheer volume and diversity of content. Recognizing these challenges, Telegram is continuing to expand its moderation framework while enhancing existing solutions – leveraging advanced software, growing its dedicated teams and fostering key partnerships to mitigate harmful content effectively, as outlined in detail below.

Telegram outlined the following features, tools, and resources it used to address harmful material and activity on its service:

- **'Content review'** – Telegram stated content on the service was reviewed 24/7 through proactive detection, user reports, 'email referrals from users and trusted organizations', and monitoring media stories. Telegram stated that it 'relies on a combination of AI / ML customized recognition tools, manual search strategies, as well as prevention of reappearance of already removed items'. Telegram stated that '[p]ublic content flagged by algorithms is processed, validated, and assigned additional priority if needed, which allows moderators to receive relevant reports and properly sort miscategorized items'.
- **'Limited content discovery'** – Telegram stated that '[b]y design, Telegram does not employ recommendation algorithms or any other form of targeted amplification'. Telegram stated that this means that bad actors cannot exploit Telegram to 'spread harmful content rapidly

[214] eSafety uses the following terms to give an impression of the extent of the indicators used by services in the table below, rather than publishing the specific indicators which could be misused:
- Minimal: A small number
- Several: A moderate number
- Multiple: A significant number.

or to reach a meaningful share of users'. Telegram said this was also true of Telegram Stories.

- **'State-of-the-art Software Solutions'** – Telegram stated that its proactive detection tools had been created by 'world-class engineers', and that it used a combination of hash and pattern matching tools and other 'state-of-the-art AI/ML models that span a wide array of technologies'. Telegram stated that models included:
    - fine-tuned self-supervised multilingual transformer-based language models;
    - fine-tuned vision transformer models;
    - multilingual transformer-based end-to-end ASR systems;
    - multimodal transformer-based models aligned on image-text datasets;
    - multilingual transformer-based large language models; and
    - custom data clustering algorithms.

  Telegram stated that 'several' of these models had been deployed by the end of the report period (29 February 2024), but it had 'since significantly expanded its use of AI and ML technologies'.
- **'Trained professionals'** – Telegram stated that its moderators are 'highly trained professionals that undergo regular quality-assurance checks including daily peer examinations'. Telegram stated that although its 'strict selection process ensures that only the most capable and suitable individuals are chosen for a moderator role', it conducted daily assessments of between 1 and 5% of all reports by randomly distributing them to moderators to calculate potential error rates. Telegram stated that 'Moderators with suboptimal error rates, or involved in systematic, gross, or material errors are replaced'. Telegram also stated that it has 'specialized moderator task groups' responsible for responding to harms such as CSAM and TVE. Telegram stated that these task forces, and its escalation processes, have 'significantly reduced the response time for handling critical reports'.
- **'Key partnerships'** – Telegram cited its collaboration with the Global Center for Combating Extremist Ideology (or, 'Etidal'). Telegram stated that between February 2022 and June 2024, Telegram's partnership with Etidal had resulted in '93,99 million pieces of content related to spreading terrorist ideologies' being removed from the service.
- Telegram stated that it was 'consistently expanding its network amongst industry leaders and international community to stay up-to-day on the latest developments and best practices related to content moderation'. Telegram also provided the following as examples of organisations that its staff regularly engaged with:
    - UK Home Office
    - Etidal

- EU Internet Forum
- Europol
- Ofcom
- UNSC Counter-Terrorism Committee Executive Directorate

Telegram stated that:

> Reports of CSAM, terrorist content and violent propaganda received from trusted organizations are processed within 1 hour.

Telegram also reiterated that, as at October 2024, it was in the process of joining the Internet Watch Foundation's safety programs to gain access to the IWF's hash lists of known CSEA material.

eSafety.gov.au