March 2025

Basic Online Safety Expectations

Mandatory transparency notices given March 2024

Key findings

Focus: Terrorist and violent extremist material and activity

Australian Government | eSafetyCommissioner

eSafety.gov.au

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). '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.

In March 2024 eSafety gave a legally enforceable notice (the Notice) to a selection of online service providers requiring them to report on measures taken to protect Australians from the risk that TVE posed to their safety and security during the report period. To help guide and align the framing of each service provider's response to the Notice, eSafety gave the following context for them to consider when answering the 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 is made or the action is done;

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 Act 1995 (Cth), which includes defences for news reports, and scientific, medical, academic or historical research, amongst others.

Supported by this context, service providers were asked to respond to questions in relation to TVE using the closest equivalent definitions in their terms of service, guidelines and policies.

Details of how service providers defined 'terrorist' and 'violent extremist' material and activity for the purposes of their terms of service, community guidelines or other equivalent service rules can be found in the full **Transparency Report**.

Key findings

Focus: Terrorist and violent extremist material and activity

The **Basic Online Safety Expectations Determination 2022** (**the Expectations**) 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.

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 certain types of illegal content, including some forms of TVE.

On 18 March 2024 eSafety gave a Notice under the *Online Safety Act 2021 (Cth)* to Google, Meta, WhatsApp, Reddit, Telegram and X Corp. It required each to detail, the steps it took to meet the Expectations by detecting and addressing online TVE on their services, for the period 1 April 2023 to 29 February 2024 (**the report period**).

Following receipt of the Notice, 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 process is ongoing.

Telegram did not comply with the Notice, as it did not respond by the deadline of 6 May 2024. eSafety subsequently received information, five months after the deadline. Telegram was given an infringement notice to deter non-compliance in the future.

This document highlights some of the key findings from responses to the Notice by Google, Meta, Reddit and WhatsApp (while WhatsApp is owned by Meta, they are considered a separate service provider for the purposes of the Basic Online Safety Expectations, so they were given a separate Notice). It also includes some information provided by Telegram after the deadline. The full transparency report contains additional information and context for these key findings. It is available at **eSafety.gov.au**.

Risks posed by particular service features

Livestreaming and video calling

Livestreamed TVE includes the broadcasting of terror attacks in live video over the internet. 'Livestreaming' was defined in the Notice as the transmission or receipt of TVE material or activity live via webcam or video to people anywhere in the world.[1] The material could be transmitted in one-on-one video calls, or video calls where one or multiple people streamed material to a group of any size.

Terrorist attacks in Christchurch[2], Buffalo[3], and Halle[4] 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.[5] In that time, approximately 200 people watched, from the terrorist's perspective, the murder of multiple people. Five years on, recordings of this footage continue to be some of the most common TVE that Australians report to eSafety.

Of the services with livestreaming and video calling functionality:

- YouTube, Facebook Live, and Instagram Live reported measures to detect livestreamed TVE
- WhatsApp and Messenger Rooms had no measures to detect livestreamed TVE in video calls, reporting that risks are mitigated by limiting the number of users on video calls – during the report period WhatsApp allowed up to 32 people on video calls, Messenger Rooms allowed up to 50 people in a video chat
- Telegram had no measures in place to detect livestreamed TVE in Channel livestreams or group video calls, reporting that its livestreaming features 'represent a generally insignificant share of the service's overall usage' and it could mitigate risks through user reporting
- Google and Meta reported that there was no mechanism to enable users that are not logged-in to YouTube or Facebook Live to make an in-service report about livestreamed TVE, even though people can access this content without logging in.

[1]Livestreaming and video calls also represent a significant risk factor for online child sexual exploitation and abuse (CSEA). eSafety has previously required online services to answer questions about their efforts to meet the Expectations by detecting and preventing livestreamed CSEA. This information can be found in our **previous transparency reports**.
[2] 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/**
[3] 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. URL: **https://ag.ny.gov/press-release/2022/attorney-general-james-and-governor-hochul-release-report-role-online-platforms**
[4] Combating Terrorism Center, 'The Halle, Germany synagogue attack and the evolution of the far-right terror threat', December 2019, URL: **https://ctc.westpoint.edu/halle-germany-synagogue-attack-evolution-far-right-terror-threat/**
[5]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/**

Generative artificial intelligence

Generative artificial intelligence[6] (**AI**) technologies offer many benefits and legitimate use cases – including for designing, implementing and supporting online trust and safety systems. However, without robust safeguards in place there are significant risks that bad actors could exploit the technology to perpetrate serious forms of online harm. For example, the ability to quickly and easily create synthetic, but highly realistic, images and videos raises significant risks that generative AI services could be misused to create vast quantities of TVE and other forms of illegal, seriously harmful content such as child sexual exploitation and abuse (CSEA) material.

Google reported it had some measures in place to stress-test and calibrate its generative AI service, **Gemini**, and to proactively prevent it from being used for harmful material and activity such as TVE and child sexual exploitation and abuse. Notwithstanding these measures Google reported that:

- It received 258 user reports about suspected AI-generated synthetic TVE by Gemini and 86 user reports of suspected AI-generated synthetic child sexual exploitation and abuse material by Gemini during the report period. In response to a follow-up question, Google said it 'was unable to confirm the number of reports confirmed to contain TVE and CSEA'.

Google also treated TVE and child sexual exploitation and abuse material differently on its Gemini service:

- Google used hash-matching[7] to scan 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).[8]
- Google used classifiers to scan text-based prompts for child sexual exploitation and abuse material, but not for TVE.

[6] Google, Meta and X Corp were asked about measures taken to safeguard their generative AI services. Meta was asked limited questions about Meta AI, as it had not been launched in Australia at the time of the report period. eSafety has not received X Corp's answers to questions about the Grok feature on X, as it applied to the AAT for review of eSafety's decision to give the Notice.
[7] Digital technology that is used to create a hash of an image or video which can then be compared against hashes of other photos to find copies of the same image or video.
[8] 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/**.

Recommender systems

Systems that use recommender algorithms can play a role in facilitating online radicalisation by progressively recommending increasingly extremist and inflammatory material to maximise engagement.[9] 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.

- Google, Meta and Reddit were asked about recommender systems in the Notice. All reported removing individual items of TVE from their services to prevent the content from being recommended to users. Google and Reddit also reported additional measures to limit the recommendation of content that may not be suitable for general audiences, whereas Meta stated:

  > "[O]ur measures are focussed on removing that content [TVE] from our services (rather than preventing its amplification)."
  > Meta, response to the Notice question asking why Meta did not have measures in place to mitigate instances of amplification of TVE on Facebook and Instagram

- Meta and Google also reported staging positive interventions to promote authoritative sources or deradicalising content on their services.
- 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'.

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

User reporting

User reporting options and complaints pathways are important safety measures because they enable users to flag and alert an online a service to specific material and activity that is illegal, harmful or otherwise in breach of its terms of service. These reporting tools are an important safety measure. There are three Basic Online Safety Expectations[10] that set out how services should be enabling user reports and the processes they should have in place to ensure these reports are properly assessed and swiftly actioned by their trust and safety systems.

- Telegram reported that it had only one reporting option for end-users to make complaints about harmful or abusive messages in Chats and Secret Chats. This option is only available when the Chat or Secret Chat is 'initiated by non-contacts and strangers'. When an end-user wishes to report a message from an account they have already added as a contact, their only option in-service is to block the user.

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.

- When a Telegram user is able to report a message in Secret Chats, which are end-to-end encrypted (E2EE), Telegram stated that it had no means of accessing messages reported by end-users. Messages reported in other parts of Telegram, such as groups and channels, were forwarded to Telegram moderators.

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.

eSafety notes that there are alternative measures that enable content moderators to review end-to-end encrypted messages that have been reported by end-users as harmful or otherwise violative. For example, WhatsApp (which is end-to-end encrypted) has processes in place that enable its moderators to receive the last five messages sent to an end-user from the account they are reporting.[11] 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.

[10] See sections 13, 14, and 15 of the **Basic Online Safety Expectations Determination 2022**.
[11] WhatsApp, 'About reporting and blocking someone on WhatsApp', accessed 15 October 2024, URL: **https://faq.whatsapp.com/414631957536067/**

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. When content such as TVE, which has the potential to cause significant harm, is reported by a user, verifying it and taking action should be done quickly to prevent ongoing or new harm.

The responses to the Notice highlighted significant differences in the time services took to consider and respond to user reports about TVE.

Median time to reach an outcome after receiving a user report about TVE[12]

More than 24 hours	Under 24 hours
Meta's Threads (59.5 hours[13])	**Meta's**[20] – **Messenger** (when E2EE enabled and not enabled[21]) (0.1 hours) – **Facebook Groups** (closed/private) (2 hours) – **Facebook Groups** (public) (2.5 hours) – **Instagram Direct** (when E2EE enabled) (Global data – 4.3 hours) – **Instagram Direct** (when E2EE not enabled) (3 hours) – **Facebook Newsfeed** (4.2 hours) – **Instagram Feed** (15.5 hours)
Reddit Public Subreddits[14] (31.3 hours[15])	
WhatsApp's[16] [17] – **Channels** (25.3 hours)[18] – **Communities** (Global data – 24.8 hours) – **Direct Messages** (including Groups (24.13 hours)[19]	
	Google's Drive (consumer version; content when it is shared) (2.9 hours[22]) **Google's YouTube** (Global data – 4.4 hours[23]) **Telegram** - **Chats and Secret Chats** (18 hours) - **Group Chats and Channels** (15 hours)

[12] Australian data unless otherwise stated.
[13] Meta noted that its figures represented 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).'
[14] Reddit reported that there were no user reports that Reddit confirmed to be terrorist content on its other services during the report period.
[15] 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 was the median time between a user report and ticket closure for reports escalated to Reddit's specialised terrorism queue.
[16] WhatsApp reported that these figures reflected 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.
[17] 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.
[18] WhatsApp reported that this information related to a total of 4 users.
[19] WhatsApp reported that information related to a total of 4 user reports.
[20] Meta noted that its figures represented 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).'
[21] Meta 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.
[22] Google reported that these figures referred to the median time taken from when a user flag was first received to when an outcome was reached.
[23] Google reported that YouTube's figures were based on data that was 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 was based on a study completed in July 2022 and that related to user flags on videos that were potentially violative of community guidelines, including guidelines related to TVE. Google reported this figure as '15 min for automated review of the flag' and 'Approx 4.4 hours for flags referred for human review'.

Trust and safety staff and language coverage

Staffing levels

In 2023, both Google and Meta announced reductions to their staffing numbers.[24] The respective announcements did not disclose how the reductions would impact the resourcing of trust and safety functions on their services.

- There was a 27.8% reduction in Meta trust and safety staff employed (other than engineers and content moderators) between 31 March 2023 and 31 Dec 2023.[25] 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) between 1 April 2023 and 29 February 2024. The number of content moderators employed by Google increased by 7.9%.

Language coverage

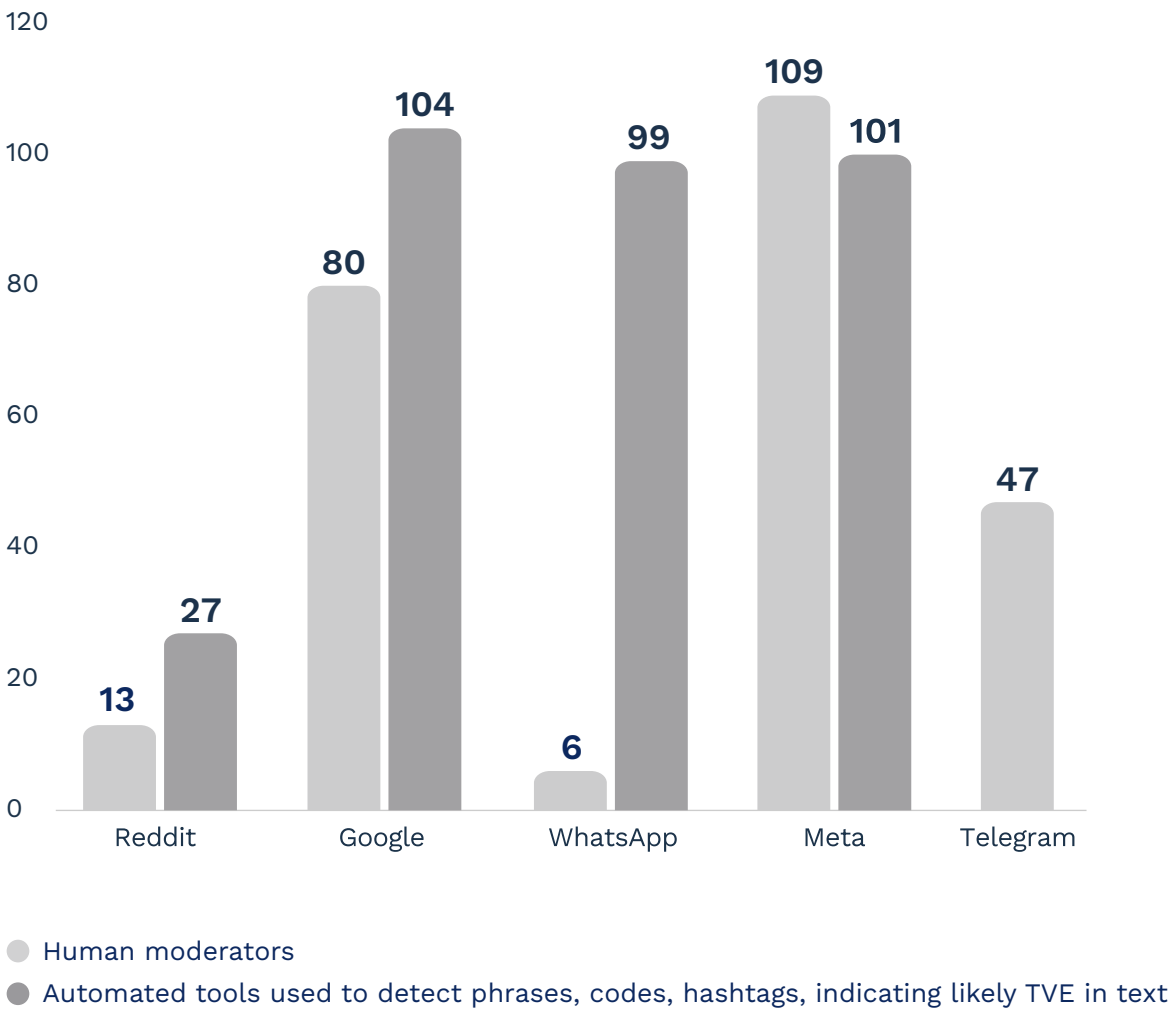
It is also important to consider the unique skillsets of the staff and contractors employed. Assessing complex, context-dependent harms requires linguistic, regional and cultural understanding. There is a risk of losing important nuance where proactive detection measures operate in a small number of languages and there is reliance on language translation tools. For this reason, it is particularly important that providers have human moderators operating in the languages of the communities to whom they offer services.

Content produced in languages other than English and Arabic is … generally more likely to remain online for longer on big tech platforms, based on our monitoring in 2022. Content in lesser-spoken languages, regional dialects, or languages used by a minority of a given platform's user base is less likely to be effectively moderated."[26]

Global Network on Extremism and Technology (GNET)

[24] 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/**.
[25] eSafety notes that over this period there was broadly no change in the number of safety engineers or content moderators.
[26] Global Network on Extremism and Technology (GNET), Trends in terrorist use of the Internet 2022, 27 Feb 2023, accessed 19 June 2024, **URL: https://gnet-research**.

Number of languages covered



The top five languages other than English spoken in Australian homes are Arabic, Cantonese, Mandarin, Vietnamese and Punjabi.[27]

- When moderating TVE, Reddit and WhatsApp human moderators covered 13 and six languages respectively and only one of the top five languages (other than English) spoken in Australian homes, despite the high use of their services in Australia.[28] In contrast, Google covered 80 languages and Meta 109 including the top five non-English languages spoken in Australian homes. Telegram's moderators covered 47 languages, but only two of the top five non-English languages spoken in Australian homes.[29]
- Google's technology on YouTube to detect phrases, codes and hashtags in text relating to TVE operated in 104 languages. Meta's technology operated in 101 languages. WhatsApp's technology operated in 99 languages. Reddit's technology operated in 27 languages across some parts of its service.
- 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.

Volunteer moderation

- Trust and safety staff at Reddit, Meta's Facebook and Telegram were not automatically informed when volunteer moderators removed an account for a TVE violation.

eSafety considers that when an offender is only banned from a specific channel or group, rather than the whole service, it can increase the risk of that offender continuing to abuse a service's terms of service related to TVE.

[27] 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)**.
[28] 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**
[29] 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.

Proactive detection and blocking

Detection

Proactive detection encompasses a broad range of interventions that services may use to discover and take action against material or activity before it is reported by an end-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.

Some people reported unintentionally seeing the video when it autoplayed on their news or video feeds. Those who watched the video included survivors of the terrorist attack as they lay in hospital, whānau of the shuhada, witnesses of the attack and ordinary people in Christchurch and around the world – adults and children alike. Almost as fast as social media platforms could remove the offensive and graphic footage, it was replaced – sometimes spliced into new video clips."[30]

Royal Commission of Inquiry into the Terrorist Attack on Christchurch Mosques on 15 March 2019

Although tools are being used to detect TVE – including images, videos and written material – they are not always applied consistently or comprehensively.

- Google used tools to detect known TVE images and videos on YouTube and on shared content in the consumer version of Drive, but not on stored content in the consumer version of Drive. Google detected new TVE videos in shared content on the consumer version of Drive but not on stored content. Google did not detect new TVE images in shared or stored content on the consumer version of Drive.

eSafety notes that Google used 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 spread of material, particularly in circumstances where edited versions of it also have the potential to go viral. For example, following the Christchurch attack Facebook identified 800 visually distinct versions of the attack video within the first days.[31]

[30] 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', 2020, accessed 18 June 2024, URL: **https://christchurchattack.royalcommission.nz/**

[31] 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/amp/?**

- Meta used hash-matching tools to detect known TVE on Messenger and Instagram Direct, but not when end-to-end encryption was enabled.

When a service is end-to-end encrypted it can limit the automated tools available to detect TVE.

- 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 enabled. Meta was reliant on user reports to detect new TVE on these services.

Notably, in 2022 Meta reported to eSafety that it was using proactive scanning tools to detect new child sexual exploitation and abuse material on Messenger and Instagram Direct (when end-to-end encryption was not enabled).

"[Meta] considers hash matching tools to be the most appropriate tool to detect TVE in private messaging threads."
Meta, 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 end-to-end encryption is not enabled.

- 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.

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.

- Telegram did not use any hash matching tools on Chats or user reports about Secret Chats. Telegram stated that this is because Telegram 'was founded on the principle of defending user privacy and their right to private communication' and that 'this commitment prioritises user privacy above all'.

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 end-to-end.

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, alternative methods also exist which could enable hash-matching tools to review content reported in end-to-end encrypted messages.

- 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.[32]

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 GIFCT or Tech Against Terrorism.

- WhatsApp rolled out Channels (which is not end-to-end encrypted) in June 2023 without implementing hash-matching for known TVE. WhatsApp reported that only during the report period did it start working on its implementation.[33]

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 retrofitted later, after the damage has been done.

> "Channels is a relatively new product" … "WhatsApp is currently working on the rollout of hash matching tools for TVE on Channels."
> WhatsApp, response to the Notice question about hash matching tools used to detect known TVE images and videos on WhatsApp

[32] 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.'
[33] WhatsApp subsequently advised eSafety that hash matching tools for TVE on Channels were deployed by May 2024.

Blocking

- While Meta did not block URLs linking to known TVE on end-to-end encrypted parts of its services, 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. The 'Safe browsing' feature is a user control, which users can turn on or off.
- While Google did block 'join-links'[34] 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.
- Telegram did not block 'join-links' and URLs to TVE across any parts of its service.

[34]A feature on some messaging services that enables end-users to forward and share access to private groups.

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 online abuse. This can take the form of multiple fake accounts, including automated accounts or bots.

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.

It is also important to consider the threat of TVE across a provider's various services and across multiple interconnected platforms. Information sharing is typically easier between a provider's own services, although the sharing of recidivism signals does also take place between providers through platforms such as ThreatExchange.

- Google, Meta, Reddit, Telegram and WhatsApp all reported having measures in place to address recidivism on their services. However, Google's Drive, Telegram and WhatsApp had minimal measures in place to address recidivism of users and groups, channels or communities.
- Instagram and Facebook mutually shared information about accounts banned for TVE and also shared information with affiliate company WhatsApp for 'severe violations of our DOI [Dangerous Organisations and Individuals[35]] and other relevant policies'. Conversely, WhatsApp did not share any information with Facebook or Instagram about accounts banned for TVE.
- Facebook and Instagram shared information about accounts banned on one service to identify accounts belonging to the same end-user on the other, but only took action to ban other identified accounts in certain specific circumstances.

eSafety considers that a key principle of Safety by Design, and the Basic Online Safety Expectations, is that information held by providers regarding abuse on one of their services is used to ensure that abuse is also not being perpetrated on their other services.

[35] 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/**

Meta's Dangerous Organisations and Individuals list

- WhatsApp's parent company, Meta, has publicly stated[36] that it maintains an internal list that designates '**Dangerous Organisations and Individuals**' that 'proclaim a violent mission or are engaged in violence' and prohibits them from its platforms.
- When asked about its access to Meta's DOI list, WhatsApp reported that it prohibits all organisations on the DOI list from using WhatsApp Channels but does not prohibit all organisations on the list from WhatsApp's private messaging.

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.

Account bans

- Google's approach on Drive was to limit bans to accounts that were 'owned or operated by a known terrorist or violent extremist organisation'.

eSafety considers that Google's approach may result in terrorists and violent extremists who are not associated with a specific organisation (such as the Christchurch attacker) evading a ban.[37]

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

[37] 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.

eSafety.gov.au