eSafetyindustry

Basic Online Safety Expectations

Addendum to transparency report and key findings related to terrorist and violent extremist material and activity – Amendments to Google LLC responses about hash-matching and user reporting

July 2025

Australian Government

eSafetyCommissioner

eSafety.gov.au

Background

On 18 March 2024 the eSafety Commissioner (**eSafety**) gave Google LLC a non-periodic reporting notice (**the Notice**) pursuant to section 56(2) of the *Online Safety Act 2021* (Cth) (**the Act**) in relation to Google Drive, Gemini and YouTube. Similar notices were issued to other providers.

The Notice focused on understanding the steps taken to address terrorist and violent extremist material and activity (**TVE**)[1] during the reporting period from 1 April 2023 to 29 February 2024.

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.

Google LLC provided a response to the Notice on 22 May 2024. This included responses to questions about hash-matching and user reporting. This information was summarised in a transparency report and key findings document published on the eSafety website on 6 March 2025.

On 9 May 2025, Google LLC provided additional information relating to its capabilities to detect known TVE images and videos. It also requested that eSafety address an omission in the transparency report and key findings document related to Google LLC's user reporting.

Each transparency report is disclosed under s 217 of the Act and is an exercise of the eSafety Commissioner's powers and functions under the Act at a point in time. As such, the initial reports cannot simply be updated. Instead, eSafety has set out the additions and clarifications in this Addendum and footnoted the original report and document to note the Addendum.

[1] 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.'

Hash-matching of known TVE

References to Google LLC that are relevant to its request for changes

- Page 12 of the key findings document states:

  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.

- Page 10 of the transparency report states:

  Google only used hash-matching to detect exact matches of TVE content, rather than edited copies.

- Page 73 of the transparency report states:

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

Table E

Parts of the 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

- Page 74 of the transparency report includes the following call-out box:

> 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.[2]

- Page 75 of the transparency report states:

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

Table F

Parts of the 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

Impact of additional information provided by Google LLC on 9 May 2025

- Google LLC stated that during the reporting period YouTube used machine learning classifiers (which work in tandem with hash-matching technology) to detect variations of known and new TVE material. eSafety advises that readers may like to note the additional information when considering Table E on page 73 and the callout box on page 74, although the Notice did not ask Google LLC whether it used technologies other than hash-matching to detect known TVE.

[2] 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 LLC stated that during the reporting period it used a proprietary partial hash-matching tool to detect partial hash-matches for videos shared in the consumer version of Drive. eSafety advises that readers may like to note the additional information, although Google LLC did not include it in its response to Notice questions about known TVE videos. eSafety advises readers that, in light of this additional information provided by Google LLC, the statements on page 10 and 74 of the transparency report and page 12 of the key findings document only apply to YouTube. For completeness, readers should be aware of this revision for Table F on page 75:

Parts of the 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 Proprietary Google hashing technology

User reporting of synthetic TVE

References to Google LLC that are relevant to its request for changes

- Page 6 of the transparency report and page 5 of the key findings document include a statement that:

  [Google LLC] received 258 user reports about suspected AI-generated synthetic TVE [being generated] by Gemini.

- Page 99 of the transparency report summarises Google LLC's response to the questions in the Notice about synthetic TVE in this way:

Table Z

Harm type	Number of user reports
TVE	258 (reviewed under Gemini's Dangerous Content policies – which includes TVE content)

Impact of additional information provided by Google LLC on 9 May 2025

- Google LLC advised that eSafety's references to reports of synthetic TVE on page 6 of the transparency report and page 5 of the key findings document should have included context, making it clear that the 'user reports' reviewed under Gemini's Dangerous Content policies were not limited to TVE content. eSafety notes the omission and advises readers that both references should be:

  [Google LLC] received 258 user reports, which were reviewed under Gemini's Dangerous Content policies (which includes, but is broader than only synthetic TVE content).

  eSafety also advises readers that Table Z should be:

Table Z

Harm type	Number of user reports
TVE	258 user reports, which were reviewed under Gemini's Dangerous Content policies (which includes, but is broader than only synthetic TVE content)