August 2025

A baseline for online safety transparency

The first regular report on child sexual exploitation and abuse, and sexual extortion

Basic Online Safety Expectations periodic reporting series

Updated 11/2025

Australian Government | eSafetyCommissioner

eSafety.gov.au

Contents

Glossary

Artificial intelligence (AI): refers to an engineered system that generates predictive outputs such as content, forecasts, recommendations, or decisions for a given set of human-defined objectives or parameters without explicit programming. AI systems are designed to operate with varying levels of automation[1].

Automated tools: Technology used to sort data into categories automatically. In the context of this report, these tools are used to support content and activity moderation actions and decisions.

Basic Online Safety Expectations: The Basic Online Safety Expectations, known as 'the Expectations' or 'BOSE', are a key element of the *Online Safety Act 2021 (Cth)* (the Act) and are set out in the *Online Safety (Basic Online Safety Expectations) Determination 2022*. They outline the Australian Government's expectations that social media, messaging and gaming service providers and other apps and websites will take reasonable steps to keep Australians safe. For more information, refer to https://www.legislation.gov.au/F2022L00062/latest/text.

Children: Individuals below the age of 18 years[2].

Child sexual exploitation and abuse (CSEA): Child sexual exploitation and abuse 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.

Child sexual abuse material: (CSAM): Material (see 'Material' glossary entry) that depicts the sexual exploitation or sexual abuse of children under 18, or person(s) who appear to be under 18.

CSEA livestreaming: The transmission or receipt of acts of sexual exploitation or abuse of children live via webcam or video to people anywhere in the world, whether or not in exchange for payment. CSEA livestreaming includes one-on-one video calls and video calls where one or multiple people stream material to a group of any size. Other Australian regulators and law enforcement agencies use other terms for this activity including live online child sexual abuse'[3].

---

[1] Australian Government, Data and Digital Government Strategy, 'Glossary', accessed 16 April 2025, URL: https://www.dataanddigital.gov.au/resources/glossary
[2] *Online Safety Act 2021* (Cth), section 5.
[3] AUSTRAC, 'Combating the sexual exploitation of children for financial gain – Financial Crime Guide December 2022', 2022, accessed 1 April 2024, URL: https://www.austrac.gov.au/business/how-comply-guidance-and-resources/guidance-resources/combating-sexual-exploitation-children-financial-gain

End-to-end encryption (E2EE): A specific method used to secure communications from one device, or 'end point', to another. E2EE transforms standard text, imagery, and audio into an unreadable format while it is still on the sender's system or device so that it can only be decrypted once it reaches the recipient's system or device.

Grooming: Predatory conduct to prepare a child for sexual abuse, including conversations and acts of contact offending.

Harmful material and/or activity: Includes material or activity that may not be unlawful, but is covered within the scope of the Act, including cyber abuse material targeting an Australian adult, cyberbullying material targeting an Australian child, and online pornography and other high-impact material. It also includes material or activity that should fall under a service provider's terms of use, policies and procedures and standards of conduct for end-users (as outlined in section 14 of the Basic Online Safety Expectations).

Hash: A numerical value that can be used to identify or match images or videos that are the same and may be referred to as a unique 'digital signature' or a 'digital fingerprint'.

Hash database: A database containing hashes that can be used to match images or videos. In the CSEA context, hash databases contain the hashes of confirmed CSEA material.

Hash-matching tools: Digital technology used to create a unique digital signature (known as a 'hash') of an image or video which is then compared against signatures (hashes) of other photos to find copies of the same image or video.

In-service user reporting: A user's ability to make reports to a service without leaving the service (for example, without being required to locate a separate webform or email address on another website).

Internet Watch Foundation (IWF): A United Kingdom-based charity that works in partnership with providers, law enforcement agencies and government to remove from the internet child sexual abuse images and videos wherever they are hosted in the world and non-photographic images hosted in the United Kingdom.

Known CSEA: Material that has been previously confirmed to contain content depicting child sexual exploitation or abuse, and which has been confirmed, hashed and stored in a hash database.

Language analysis tools: Technologies using methods to assign a probability that text or conversations involve certain behaviour, such as grooming a child for sexual abuse. This may involve artificial intelligence programming.

Lantern: A cross-platform information sharing program led by the Tech Coalition.[4]

Machine Learning: A type of artificial intelligence that gives computers the ability to learn without explicitly being programmed, which can be applied to new data for prediction or decision-making purposes.

Machine learning classifiers: A classifier is an algorithm that automatically orders or categorises data into one or more of a set of 'classes.' They are rules that map input data into predefined categories.

Material: Defined in section 5 of the Act to mean material, whether in the form of text, data, speech, music or other sounds, visual images (moving or otherwise), of any other form or any combination of forms. In the context of CSEA, 'material' typically refers primarily to images or video content.

NCII: Non-consensual intimate image. For the purposes of the Act, an intimate image can include a still image or moving image (such as a video) depicting, or appearing to depict private parts (such as the genital area), private activity (such as a person in a state of undress, showering or engaged in a sexual act of a kind not ordinarily done in public) or a person without attire of religious of cultural significance, in circumstances in which an ordinary reasonable person would reasonably expect to be afforded privacy (section 15 of the Act) that is shared without the consent of the person shown.

NCMEC: The United States based National Center for Missing and Exploited Children. NCMEC hosts databases of confirmed CSEA hashes, which enable providers to detect when this content is uploaded to their services.

New CSEA: Material that has not been previously confirmed to contain content depicting child sexual exploitation or abuse, and which has not previously been hashed and stored in a hash database. Also known as 'first generation CSEA'.

Notice: For the purpose of this report, a Notice is a periodic notice given to a provider under section 49(2) of the Act on 22 July 2024.

Recidivism: In the context of this report, banned or suspended users re-registering to a service with new details to continue perpetrating online abuse. Multiple fake or imposter accounts may be created by a human user or by an automated bot programmed by a human.

Recommender systems: Algorithm-based programming that makes personalised content suggestions to users based on a range of factors.

[4] For more information, refer to: https://www.technologycoalition.org/newsroom/announcing-lantern

Report period: When providers receive a reporting Notice from eSafety they are required to prepare a report about the extent to which they complied with the Expectations during a specified period. This period is referred to as the 'report period'. The report period for this set of Notices is 15 June 2024 to 15 December 2024. Information provided reflects this period, unless stated otherwise.

Sexual extortion: Also known as 'sextortion', is a crime involving online blackmail, where victims are tricked into sending intimate images or videos of themselves to someone who then threatens to share the images or video unless demands are met, usually for payment. Deepfakes are increasingly being used to perpetrate this crime regardless of whether the victim ever sent intimate images or videos to someone else. There has been a substantial growth in sexual extortion, targeting teenage males in particular. Sexual extortion of someone under 18-years-old is a form of child sexual exploitation and abuse.

Stop NCII: An organisation that provides hash lists to providers to scan images or videos on their services for non-consensual intimate image abuse. Refer to https://stopncii.org/.

Take It Down: A NCMEC initiative that provides hash lists to providers to scan images and videos on their services for child sexual exploitation material. For more information, refer to https://takeitdown.ncmec.org/

The Act: The Online Safety Act 2021 (Cth).

Unlawful material and/or activity: Includes material or activity that is not permitted under an Act of the Commonwealth of Australia. For the purposes of the Determination, the term 'unlawful' refers to illegal material or activity dealt with under the Act and other unlawful material or activities that may have a negative impact on the online safety of Australians.

URL: Uniform Resource Locator, colloquially known as the address of a web page.

URL database: For the purposes of this report, databases of URLs linking to known CSEA material kept to enable providers to prevent use of those URLs on their services. An example of a URL database is the one provided by the Internet Watch Foundation.[5]

Video-calling service / video chat service: A service that facilitates two-way audio and video communication between two or more devices equipped with cameras and screens, allowing users to see each other as they talk. Video-calling can be one-on-one, but it can also involve multiple participants, for example in video conferencing services.

[5] Internet Watch Foundation URL List, accessed 9 October 2023

Introduction

The purpose of this periodic report series: shining a light on online industry action against child sexual exploitation and abuse

This report, the first of four to track industry progress over two years, examines how major technology companies are responding to the persistent challenge of addressing **child sexual exploitation and abuse (CSEA)**, including **grooming** and **sexual extortion**, as well as sexual extortion of adults, on their services.

Since our inception 10 years ago, eSafety has identified CSEA as one of the most acute and enduring forms of online harm we encounter. INHOPE, a global network of hotlines working to eliminate CSEA from the internet, reported that in 2024 it received record numbers of reports of previously unseen materials, reflecting a 34.8% increase compared to 2023.[6] However, this only represents a piece of this global issue. For example, **the National Center for Missing and Exploited Children (NCMEC)'s** CyberTipline is the United States' centralised reporting system for the exploitation of **children** online. In 2024 alone, more than 20.3 million incidents of suspected **child sexual abuse material** were reported to NCMEC. As these organisations receive reports globally, these sources may be reflective of Australia's experience of exploitation of children online. Similarly, Research by the Global Partnership and Fund to End Violence Against Children states that at 'any given time, an estimated 750,000 people are looking to connect with children for sexual purposes online'.[7] Research by the Childlight Global Child Safety Institute estimates that around one in eight children globally (more than 300 million children) each year are victims of sexual abuse and exploitation online. Additionally, in 2024 approximately one in eight children worldwide is estimated to have experienced online solicitation, including non-consensual sexting, unwanted sexual questions or unwanted sexual act requests by adults or other youths in the past year alone.[8]

Australia's *Online Safety Act 2021* (Cth) (the Act) aims to address child sexual exploitation and abuse online, among other harms.

---

[6] INHOPE, 2024, 'Annual report 2024', accessed 28 May 2025, URL: https://inhope.org/media/site/41f00cc3d9-1743600476/inhope-annual-report-2024.pdf

[7] End Violence Against Children, 2024, 'Issue and response', accessed 26 June 2024, URL: https://www.end-violence.org/node/7939

[8] UNSW Media, Childlight Global Child Safety Institute, 2024, 'More than 300 million child victims of online sexual abuse globally: report', accessed 26 June 2024, URL: https://www.unsw.edu.au/newsroom/news/2024/05/more-than-300-million-child-victims-of-online-sexual-abuse-globally-report

On 22 July 2024, the eSafety Commissioner gave periodic **notices** under section 49(2) of **the Act** to eight significant providers: Apple, Discord, Google, Meta, Microsoft, Snap, Skype,[9] and WhatsApp.

The periodic notices require these service providers to report every six months for a two-year period on how they are implementing the **Basic Online Safety Expectations** (Expectations) established by the Australian Government. Providers were asked about steps they are taking to meet several Expectations, including Expectations to ensure that end users are able to use the service in a safe manner, the services have mechanisms for end users to report and make complaints, and that the providers take reasonable steps regarding their **end-to-end encrypted (E2EE)** services, **artificial intelligence** capabilities, **recommender systems** and to.[10]

This transparency series builds upon eSafety's previous reports published in December 2022 and October 2023, which highlighted significant variations in safety approaches and concerning protection gaps in relation to CSEA. In giving the periodic notices in 2024, the eSafety Commissioner noted that the companies were chosen partly based on answers many of them provided to eSafety previously in response to non-periodic notices, exposing a range of safety concerns when it came to protecting children from abuse, as well as other considerations listed under section 49(5) of the Act.

Their responses, covering the period 15 June to 15 December 2024, reveal some progress but also concerning gaps in protection measures.

Connecting the patterns of exploitation

The periodic notices also include questions about sexual extortion – a crime affecting both children and adults – to address the interconnected nature of these harms.

Sexual exploitation and abuse can take different forms, including extortion. For example, it could include any or all of the following steps:

- Connecting with a child in an online environment and luring them into an inappropriate online chat (this is called 'grooming' them).
- Coercing them into taking and sending nude photos of themselves (creating online CSEA).
- Threatening to share the photos with the child's family unless they agree to being sexual in a video call (sexual extortion).
- Selling a recording of the video call with the child via an online forum (a form of CSEA) for the sexual gratification of other consumers (whose purchase of the CSEA is also illegal).

[9] Microsoft announced the deprecation of Skype from 5 May 2025: Microsoft, 2025 'The next chapter: Moving from Skype to Microsoft Teams', accessed 3 March 2025, URL: https://www.microsoft.com/en-us/microsoft-365/blog/2025/02/28/the-next-chapter-moving-from-skype-to-microsoft-teams/

[10] Providers were asked about Expectations 6, 8, 8A, 8B, 9, 10, 11, 12, 13, 14, 15 and 17, as relevant to their services.

eSafety's investigations reveal concerning patterns where tactics used in adult sexual extortion are increasingly being adapted to target children, while grooming techniques continue to evolve.

By examining sexual extortion targeting both children and adults, we gain crucial insights into how platforms can develop comprehensive safety frameworks that protect all users from exploitation, regardless of age.

Anticipating new challenges

As generative AI technologies rapidly evolve, bringing both new opportunities and risks, our report series extends to examine how companies are addressing emerging threats in this space. These powerful technologies can both amplify existing CSEA risks through the creation of synthetic CSEA and introduce entirely new vectors for exploitation.

In May 2025, NCMEC reported that its CyberTipline had seen 'a 1,325% increase in reports involving Generative AI, going from 4,700 last year to 67,000 reports in 2024'[11]

The U.S. Department of Homeland Security reported that there are a variety of ways that offenders can use AI to create CSEA. Offenders can use AI to:

- edit an image of a child to make it appear as though the child is nude or engaged in sexual acts.
- create an image of a child being sexually abused via text prompts.
- manufacture images of children being abused who look like real people but are fabricated.
- teach other offenders how to engage with children online (such as how to groom children online).
- edit previously created and shared content to create new sexual abuse material, revictimising CSEA victims.[12]

Collecting data indicating how platforms are proactively managing these emerging risks provides crucial understanding for developing effective regulatory frameworks that can keep pace with technological innovation.

What this report includes

This report summarises the information that eSafety received from responses to the Notice. 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

[11] National Centre for Missing & Exploited Children (NCMEC), 2025, 'CyberTipline 2024 Report', accessed 26 May 2025, URL: https://www.missingkids.org/gethelpnow/cybertipline/cybertiplinedata

[12] U.S. Department of Homeland Security, n.d, 'Artificial intelligence and combatting online child sexual exploitation and abuse', accessed 27 May 2025, URL: https://www.dhs.gov/sites/default/files/2024-09/24_0920_k2p_genai-bulletin.pdf

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**.

The following points should also be noted:

- eSafety has not run technical tests or research to test the veracity of every response given by providers in response to the Notice, however, our comments throughout this summary are based on our insights and experience working with each of these services over the past decade. 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 15 June to 15 December 2024, inclusive, or other periods within this timeframe as specified.
- All data is global, unless otherwise stated.
- Bolded terms are defined in the glossary of this report, unless otherwise stated.

Beyond compliance: The responsibility to act now

While acknowledging that each platform has unique architectures, business models, and user bases – all requiring tailored safety approaches – transparency creates valuable opportunities for cross-industry learning and innovation. Our findings show that even within the same parent companies, detection tools and safety measures vary across services. This demonstrates that a common baseline of protection remains elusive.

By enabling public scrutiny, this report series serves two critical functions: incentivising meaningful safety improvements across the technology sector, and holding companies accountable for protecting their most vulnerable users.

The Expectations work alongside Australia's Online Safety Phase 1 Codes and Standards.[13] These Codes and Standards place mandatory and enforceable obligations on relevant participants in the online industry requiring them to take action to combat class 1 material such as CSEA and pro-terror content. In some cases, questions asked of industry under periodic notices about their compliance with the Expectations may also be relevant to their compliance with Phase 1

[13] eSafety, 2025, 'Register of industry codes and industry standards for online safety', accessed 13 June 2025, URL: https://www.esafety.gov.au/industry/codes/register-online-industry-codes-standards

Codes and Standards. There are a range of enforcement mechanisms for non-compliance with the Phase 1 Codes and Standards, including civil penalties.

Industry leaders, policymakers, civil society groups, researchers and law enforcement can use this report as a tool for change. Transparency is not just about disclosure – it is about driving real safety improvements.

Through this report series, eSafety will closely monitor the responsiveness of technology companies to the need to address online harms and provide meaningful and fulsome information about the steps they are taking. eSafety also has the power to give and publish statements about a provider's compliance with the Expectations.[14] Information obtained and published through these notices will provide a useful evidence base for determining where providers may be non-compliant.

In publishing this report, eSafety expects industry to learn from each other and make concerted efforts to uplift their safety, and take a best practice approach to meeting the Australian Government's Expectations.

[14] eSafety refers to these as 'Statements of Non-Compliance' and 'Statements of Compliance'. More information can be found in Part 3 of eSafety's Regulatory Guidance on the BOSE, available at: https://www.esafety.gov.au/industry/regulatory-guidance#basic-online-safety-expectations

Proactively detecting CSEA livestreaming

CSEA livestreaming is the transmission or receipt of acts of sexual exploitation or abuse of children live via webcam or video to people anywhere in the world, whether or not in exchange for payment. CSEA livestreaming includes one-on-one video calls and video calls where one or multiple people stream CSEA material to a group of any size.

Despite the availability of technology to help detect child sexual exploitation and abuse livestreaming or video calls, no providers were using it on all parts of their service(s).

Providers not using tools to proactively detect CSEA livestreaming

Apple did not use tools to detect CSEA livestreaming on FaceTime.

Discord did not use tools to detect CSEA livestreaming on Go Live or Video Calls (voice and video calls became end-to-end encrypted in September 2024).

Google did not use tools to detect CSEA livestreaming on Google Meet, but did use tools on YouTube.

Meta did not use tools to detect CSEA livestreaming on Facebook Messenger, but did use tools on Facebook Live.

Microsoft did not use tools to detect CSEA livestreaming on Teams*

Skype did not use tools to detect CSEA livestreaming.

Snap did not use tools to detect CSEA livestreaming in Snapchat Video Chats.

WhatsApp did not use tools to detect CSEA livestreaming in video calls.

*Sentence corrected 11 November 2025

Proactively detecting new CSEA material

Tools can be deployed on services to detect the sharing of CSEA material when it is first created, and before it has been verified and included in a database. These tools help stop the spread of CSEA and alert providers to users who are engaging in this illegal activity.

While most services were using tools to detect new CSEA, some were not.

Providers not using tools to proactively detect new CSEA

Apple did not use tools to detect new CSEA on iCloud or iCloud email.

Microsoft did not use tools to detect new CSEA material on OneDrive or Outlook.

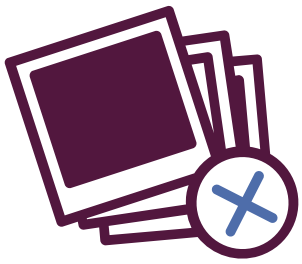
Skype did not use tools to detect new CSEA.

Snap did not use tools to detect new CSEA on material on Private Stories and Chats, which are private surfaces, unless that material was reported to Snap.

Google did not use tools to detect new CSEA on Google Meet, Google Chat, Google Messages or Gmail. Although Google did use tools to detect new CSEA material on Google Drive through classifiers, Google only did this after an account was flagged as 'suspicious'.

Ongoing safety issues

eSafety has previously reported on the inaction of these providers in detecting new CSEA on these services in our 2022 and 2023 transparency reports on CSEA.



Of the **28 million images** reported to NCMEC in 2024, 44% were new/unique images. Of the **33.1 million videos** reported to NCMEC, 25% were new/unique. Source: https://www.missingkids.org/gethelpnow/cybertipline/cybertiplinedata

INHOPE reported that the reports it received of new records in 2024 (929,733) reflected a **34.8% increase** compared to 2023. This 'surpassed the peak levels seen in 2021 (760,054) by 22.4%, demonstrating a significant rise in the identification of previously unseen materials' Source: https://inhope.org/EN/articles/inhope-annual-report-2024

User reporting to identify CSEA material and activity

Providing clear and readily identifiable reporting mechanisms is a core expectation of the Basic Online Safety Expectations. User reporting can prompt service providers to remove CSEA in a timely manner and to report to appropriate authorities. It is a critical safety intervention for all services, but especially for those that have end-to-end encryption which can limit the use of some proactive detection tools.

Since 2022, when eSafety first gave notices to providers, three services had implemented end-to-end encryption: Google Messages, Discord and Facebook Messenger.

While most services provided user reporting options and stated they responded to user reports in a reasonable amount of time, there were some providers who took much longer.

Providers with safety deficiencies in user reporting

Apple did not provide in-service reporting for CSEA on iCloud email, iCloud or FaceTime (E2EE). Apple was required to provide the number of CSEA reports it made globally or in Australia, as well as the median time to respond to those reports. Apple did not provide a response to this question.

Google did not provide in-service reporting for Gmail or Messages (E2EE). YouTube did not allow users to make reports without logging in, nor did it have a specific reporting category for CSEA, though it did have 'child abuse' and 'sexual content' reporting categories under which CSEA could be reported.

Discord id not provide in-service user reporting for CSEA on Go Live or E2EE video calls.

Whatsapp did not have a specific CSEA reporting category (WhatsApp Messages were E2EE).

Ongoing safety issues

eSafety has previously reported on the lack of in-service reporting for CSEA on Apple services, Gmail and Google Messages, and Discord's livestreams and audio, in our 2022 and 2023 transparency reports on CSEA.

NCMEC's analysis has shown that further implementation of E2EE by providers has contributed to a **decrease** in reports. This is because it is harder for providers to implement tools to detect CSEA on E2EE services. Source: https://www.missingkids.org/gethelpnow/cybertipline/cybertiplinedata

Threads' human moderators responded to reports fastest, taking a median time of **33 minutes** globally and **40 minutes** for Australian users. WhatsApp was slowest, with an estimated median time of **1,621 minutes** (27 hours) globally, and **1,594 minutes** (26.5 hours) for Australian users.

Proactively detecting known CSEA material

Hash-matching is a widely available tool that providers can easily deploy to detect the sharing of **known CSEA** material. It is a key safety measure.

Hash-matching is a form of digital 'fingerprinting' that allows copies of previously identified CSEA images and videos to be detected with very high levels of accuracy. Hash matching allows this content to be quickly removed in a privacy preserving way and flags that the service needs to review or action the relevant account(s).

While most providers were using hash-matching on their services (other than end-to-end encrypted services or parts of services) not all services were using this tool.

Providers not using hash-matching to proactively detect known CSEA

Apple did not use hash-matching to detect known CSEA images on iCloud, or known CSEA videos on iCloud or iCloud email.

Discord did not use hash-matching to detect known CSEA videos.

Google did not use hash-matching to detect known CSEA images on Google Messages nor to detect known CSEA videos on Gmail, Google Chat or Google Messages

Microsoft did not use hash-matching to detect known CSEA images or videos stored on OneDrive nor to detect known CSEA videos on Outlook (Microsoft only used hash-matching to detect when known CSEA images or videos were shared).

Ongoing safety issues

eSafety has previously reported on the lack of hash-matching tools used for known CSEA images and/or videos on all of these services in our 2022 and 2023 transparency reports on CSEA.

In 2024, NCMEC received 20.5 million reports of suspected child sexual exploitation. The reports to NCMEC contained **'62.9 million images, videos and other files related to the child sexual exploitation incident being reported.'** Source: https://www.missingkids.org/gethelpnow/cybertipline/cybertiplinedata

In 2024, NCMEC's Take It Down service (a free service that helps victims-survivors of online CSAM remove from the internet nude, partially nude or sexually explicit photos and videos taken before they turned 18) received more than **83,000 submissions** including more than 166,000 hashes. Source: https://www.missingkids.org/gethelpnow/cybertipline/cybertiplinedata

Proactively detecting other CSEA activity

Sexual extortion is a form of blackmail where someone threatens to share a nude or sexual image or video unless the person targeted gives in to their demands (usually for money or additional intimate material). Sexual extortion of someone under the age of 18 is a form of CSEA activity.

There are tools, such as language analysis tools, that services can use to detect sexual extortion and stop this criminal activity, but not all of them were using these tools and not all tools were calibrated to keep users of all ages safe.

Providers not using language analysis tools to proactively detect sexual extortion

Apple did not use language analysis tools to detect sexual extortion of adults on iMessage, Facetime, or iCloud email.

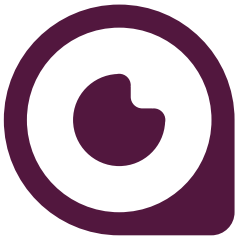
Discord did not use language analysis tools to detect sexual extortion of adults on any part of its service. Discord only used language analysis tools to detect sexual extortion of children in direct messages (and not on any other part of the service, including discoverable community servers, community servers or friend servers).

Google did not use language analysis tools to detect sexual extortion of adults or children on Google Meet, Google Chat or Google Messages.

Microsoft Teams did not use language analysis tools to detect sexual extortion of adults or children.

Skype did not use language analysis tools to detect sexual extortion of adults or children.

Snap's language analysis tools to detect sexual extortion did not operate in any of the most common languages spoken in Australian homes other than English (Mandarin, Arabic, Vietnamese, Cantonese or Punjabi). These tools were only used on user reports, not proactively.



'In 2024, NCMEC received nearly **100 reports** of financial sextortion a day and since 2021, NCMEC is aware of more than **three dozen teenage boys** who have taken their lives as a result of being victimized by this crime.' Source: https://www.missingkids.org/gethelpnow/cybertipline/cybertiplinedata

How to read this report

Providers were required to report on the measures they were taking during the reporting period to address CSEA on their services. 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.

Where providers reported on information that addressed the same or similar issues, eSafety has compiled that information in summary tables or graphs. Setting it out in this way allows easy comparison, which enables a fuller understanding of the differences in how the services operated.

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 and graphs 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.

Responses by issue

Proactive detection of CSEA

CSEA material and activity can be detected in two main ways: proactive detection by use of tools and technologies, and by user reporting.

Proactive detection encompasses a broad range of interventions that service providers may take to discover and take action against CSEA on a service before it is reported by a user. These interventions typically involve the use of technologies and tools to automatically detect material or activity that is prohibited by a service's terms of service. Further information about these tools and technologies is set out in this section.

This section looks at how CSEA material (both new and known) and CSEA activity (including livestreaming, grooming and sexual extortion) are generally detected, as well as specific detection issues on end-to-end encrypted services.

Proportion of CSEA material flagged by automated tools versus reported by users

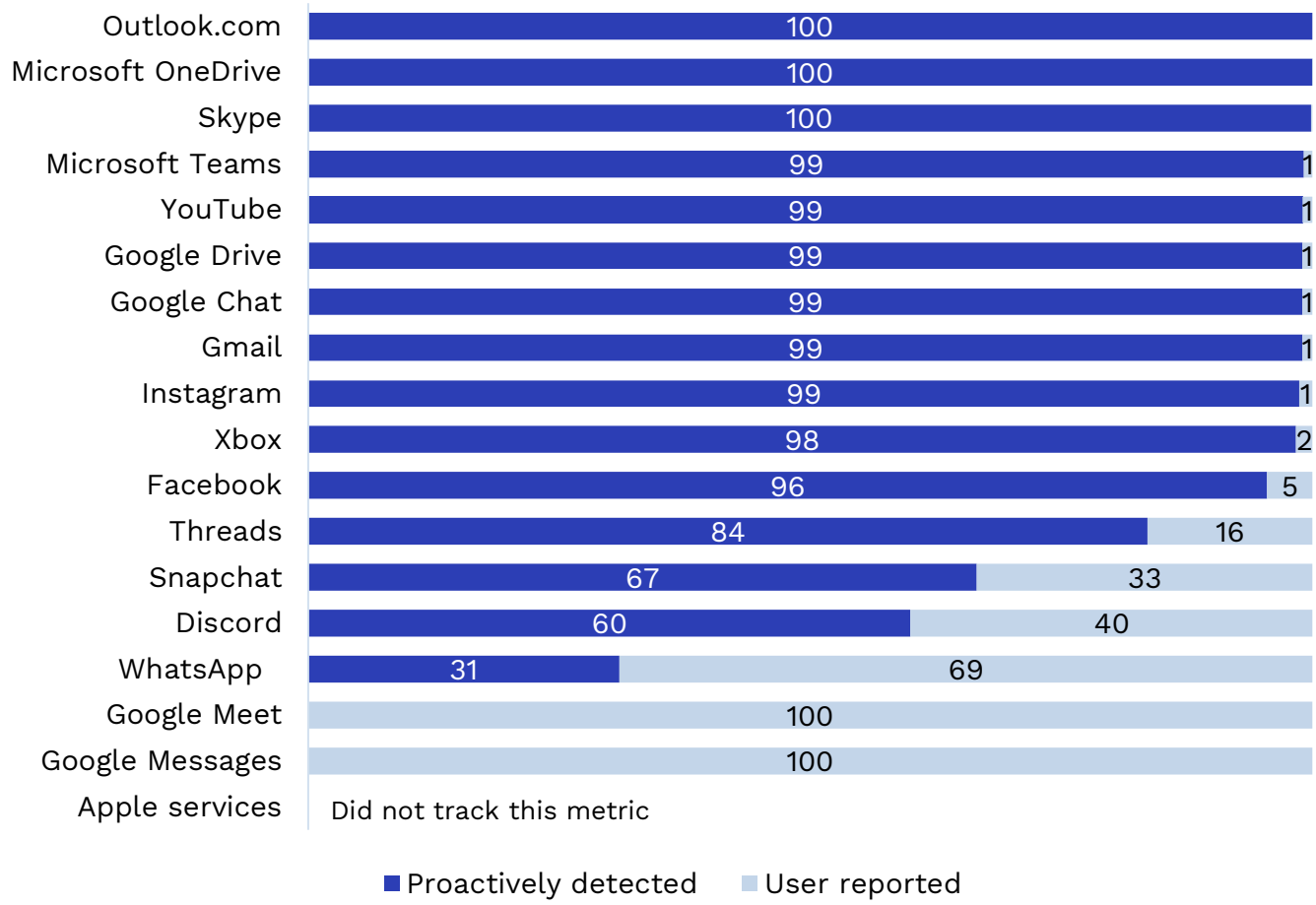
Providers were asked questions about the percentage of CSEA material that was removed following proactive detection versus user reports. This information gives useful insights into

how providers are identifying and removing CSEA, and in particular, how reliant they are on users to report harm.

The information also gives some insights into how effective proactive detection tools are for a service. However, it should not be assumed that if a service had a higher percentage of CSEA material reported by users than any other service it was an indication that there was more CSEA on that service. It could simply indicate that user reporting options were easier to find and use on that service.

Consistent with other providers, Apple's notice required it to report the percentage of CSEA material that was detected proactively versus reported by users on its services during the **report period**. Apple stated that it was not able to provide this information as it was not tracking it during the relevant period. Apple stated that it was setting up a process to measure the proportion of such material detected proactively versus reported to Apple by users.

Figure 1: CSEA material that was detected proactively versus reported by users (%)



Note 1: Data has been rounded to full numbers. Microsoft OneDrive and Skype recorded 0.13% or less for user reported CSEA.

Note 2: Apple stated that it did not scan for CSEA material on iMessage, FaceTime or iCloud. Apple stated that it was using perceptual hashing technology to detect known CSEA images on iCloud email, thus, this technology would

likely proactively detect at least some CSEA on iCloud email. Apple also stated that Communication Safety may have detected some CSEA on FaceTime and iMessage, though Apple was unable to provide the proportion of CSEA material it detected.

Note 3: Discord calculated this metric by evaluating messages that were proactively deleted for CSEA, then dividing that number by the total number of CSEA messages removed from Discord during the report period.

Note 4: Google (YouTube, Google Drive, Google Meet, Google Chat, Google Messages, Gmail) stated that it understood 'detected proactively' to mean any suspected CSEA detected by use of either hash-matching or machine learning classifiers technology. Google stated that no automated tools were used on Meet and Google Messages.

Note 5: Meta (Facebook, Instagram, Threads) stated that the proportions were calculated by identifying all material that violated its child sexual exploitation, abuse and nudity policy and calculating: 1) the percentage of such material that was removed before a user reported it to Meta; and 2) the percentage of such material that was removed after a user reported it to Meta. Meta stated that it was unable to calculate the percentage of CSEA that was detected proactively versus reported by users on Instagram Direct or Facebook Messenger because it did not have data which linked enforcement action on these services to a specific user report. If a user reported a message on these services that violated its policies, Meta removed the message attachment (for example a photo or video) from the message thread (or otherwise took enforcement action against the account) but did not remove the message itself.

Note 6: Microsoft (OneDrive, Outlook.com, Teams, Xbox) stated that it calculated the percentage of material that was proactively detected versus reported by calculating the number of confirmed CSEA material that was reported by users against the total number of confirmed CSEA material; then calculating the percentage of confirmed CSEA material that was identified proactively against the total number of confirmed CSEA material.

Note 7: Skype calculated the percentage of material that was proactively detected versus reported by calculating the percentage of confirmed CSEA material that was reported by users against the total number of confirmed CSEA material; then calculating the percentage of confirmed CSEA material that was identified proactively (such as, using scanning technologies) against the total number of confirmed CSEA material.

Note 8: Snap (Snapchat) stated that this metric reflected the ratio between the number of enforcements for proactively detected CSEA and the total count of enforcements for CSEA in the reporting interval (irrespective of whether these enforcements were taken following proactive detection or user reporting). Snap also stated that it sampled Stories to assess percentage views of illegal or otherwise violating content and determined that the prevalence of CSEA Stories was extremely low (Snap also noted this in its EU DSA Risk and Mitigation Assessment Report).

Note 9: WhatsApp stated that it calculated this metric by identifying all accounts that were banned for CSEA-related violations and calculating the percentage of such accounts that did not have a user report against them in the 30 days prior to enforcement ('proactively detected') and the percentage of such accounts that did have a user report against them in the 30 days prior to enforcement ('reported by users').

Proactively detecting known CSEA material

Hash-matching is a form of digital 'fingerprinting' that allows copies of previously identified CSEA to be detected at very high levels of accuracy. PhotoDNA is an example of a hash-matching tool, developed by Microsoft and Dartmouth University in 2009, and it reportedly has an error rate of approximately 1 in 50 billion.[15] 'Hashes' of previously identified CSEA images are stored in databases operated by expert organisations, and internally by companies, after they have been verified by expert analysts. For example, the U.S.-based NCMEC triple vets that images and videos contain child sexual abuse before they are added to their global **hash database**, which can then be used by providers to detect matches (or near matches) uploaded to their services.[16] In 2024, NCMEC received 20.5 million reports of suspected child sexual

[15] Farid, H., 2019, 'House Committee on Energy and Commerce Fostering a Healthier Internet to Protect Consumer', accessed 17 July 2025, URL: https://www.congress.gov/116/meeting/house/110075/witnesses/HHRG-116-IF16-Wstate-FaridH-20191016.pdf

[16] National Center for Missing and Exploited Children, 2023, 'United States Senate Committee on the Judiciary 'Protecting Our Children Online' February 14, 2023 Testimony of Michelle DeLaune, President and CEO National Center for Missing & Exploited Children', accessed 17 July 2025, URL: https://www.missingkids.org/blog/2023/congress-commits-to-protecting-children-online

exploitation. The reports to NCMEC contained '62.9 million images, videos and other files related to the child sexual exploitation incident being reported.'[17]

Providers were asked about their use of hash-matching tools to detect known CSEA images.

Key insights:

- Apple did not use hash-matching tools to detect known CSEA images on iCloud (which was opt-in end-to-end encrypted) and did not use hash-matching tools to detect known CSEA videos on iCloud or iCloud email. For iMessage and FaceTime (which were end-to-end encrypted) Apple only used Communication Safety, Apple's safety intervention to identify images or videos that likely contain nudity, as a means of 'detecting' CSEA.
- Discord did not use hash-matching tools for known CSEA videos on any part of the service (despite using hash-matching tools for known images and tools to detect **new CSEA** material).
- Google did not use hash-matching tools to detect known CSEA images on Google Messages (which was end-to-end encrypted), nor to detect known CSEA videos on Google Chat, Google Messages or Gmail.
- Microsoft did not use hash-matching tools for known CSEA images stored on OneDrive[18], nor did it use hash-matching tools to detect known videos within content stored on OneDrive or on Outlook.

Compared with previous responses[19]

eSafety previously reported on gaps in the use of hash-matching tools to detect known CSEA images and/or videos on the services mentioned in the Key Insights in our 2022 and 2023 transparency reports on CSEA.

While many gaps remain, eSafety has observed positive changes on some services:

[17] National Centre for Missing & Exploited Children (NCMEC), 2025, 'CyberTipline 2024 Report', accessed 26 May 2025, URL: https://www.missingkids.org/gethelpnow/cybertipline/cybertiplinedata

[18] Microsoft stated that it intends to begin hash-matching for known CSEA and terrorist or violent extremist content in limited jurisdictions including Australia.

[19] Non-periodic notices were given in 2022 to Apple, Meta, Microsoft, Skype, Snap and WhatsApp, and in 2023 to Discord and Google. Note that the questions asked and the reporting timeframes differed between notice rounds and between providers. These comparisons serve as a guide to how service responses may have changed since the last notice they received. Future comparisons will be observed through the following three periodic notice responses from providers.

- Discord applied hash-matching tools to detect known CSEA images on more parts of its service than it did in 2023. In addition, the number of tools that Discord used to detect known CSEA increased since 2023.
- Microsoft extended its application of hash-matching tools to detect known CSEA videos to content that is shared on OneDrive.
- WhatsApp increased the number of hash-matching tools it used on its service.

Table 1: Hash-matching tools used to detect known CSEA images

Provider	Service or part of service	Name of tools used	Were the tools used on all or only parts of the service?	Were the tools used on all material by default, or only in certain circumstances?
Apple	iCloud email	PhotoDNA	All parts (outbound and inbound mail)	All material by default
Discord	Discoverable community servers Community servers Friend servers Direct messages	PhotoDNA CLIP PDQ hashes	All messages sent in Discord servers and via direct or group messages, banners, server avatars and user avatars	All material by default
Google	YouTube	CSAI Match	All videos	All material by default
	Drive (material that is stored) Drive (material that is shared) Chat Gmail	PhotoDNA SHA256 Internal proprietary technologies	Drive: All stored and shared materials, including those extracted from uploaded files. Gmail and Chat: All images extracted from uploaded files.	All material by default
	Gemini	PhotoDNA SHA256 Internal proprietary technologies	All uploaded images, including those extracted from uploaded files	All material by default
Meta	Facebook Threads	PhotoDNA SimSearchNet++ (SSN++) PDQ	All	All material by default
	Facebook Messenger	PhotoDNA	On parts of the service where end-to-end	All material by default

		PDQ	encryption was not enabled (including user reports from end-to-end encrypted threads)	
	Instagram	PhotoDNA SimSearchNet++ PDQ	Instagram: All Instagram Direct: On parts of the service where end-to-end encryption was not enabled (including user reports from end-to-end encrypted threads)	Instagram: All material by default Instagram Direct: All material by default
Microsoft	OneDrive (material when it is shared) [20]	PhotoDNA MD5	All	All material by default
	Outlook.com	PhotoDNA MD5	Outbound email attachments sent from North America that could go to Australian end-users[21]	All material by default
	Teams	PhotoDNA MD5	Used on image and video attachments in chat	All material by default
	Xbox	PhotoDNA MD5	In all features where imagery could be uploaded and shared	All material by default
Skype	Skype	PhotoDNA MD5	All images sent on Skype including profile pictures (except in Skype Private Conversations[22])	All material by default
Snap	Spotlight Discover Stories Chat	PhotoDNA	Material uploaded from a user's camera roll	Material uploaded from a user's camera roll
WhatsApp	WhatsApp (other than Channels)	PhotoDNA SimSearchNet++	Material reported by users, group picture, user profile picture and community picture	All material by default
	WhatsApp Channels	PhotoDNA SimSearchNet++	Channel profile picture, Channel posts	All material by default

[20] Microsoft stated that OneDrive intends to begin using hash-matching tools to detect known CSEA images and terrorist or violent extremist content on new material that is stored on the service in limited jurisdictions, including Australia.
[21] Microsoft stated that Outlook would be investing resources into expanding the scanning for all outbound emails worldwide.
[22] Skype Private Conversations uses end-to-end encryption.

Note 1: Apple's iMessage, Facetime and iCloud; Google Messages; Microsoft's OneDrive (material that is stored); and Snap (in Snaps) did not use hash-matching to detect known CSEA images and are therefore not included in the table.

Note 2: Snap stated that in the reporting interval, it used PhotoDNA to proactively scan material uploaded from a user's camera roll to Snapchat by default. This was applied regardless of where in Snapchat a user was seeking to upload an image (for example, whether to Spotlight, Discover, Stories or Chat). Snap stated that it used this technology to detect known CSEA material on all surfaces of Snapchat where there could reasonably be a hash match; Snap did not use hash-matching to detect known CSEA images or videos on Snaps, which are novel photos and videos taken using the Snapchat app camera.

Note 3: Microsoft stated that OneDrive intends to begin using hash-matching tools to detect known CSEA images and terrorist or violent extremist content on new material that is stored on the service in limited jurisdictions, including Australia.

Table 2: Hash-matching tools used to detect known CSEA videos

Provider	Service	Name of matching tools	Were the tools used on all or only parts of the service?	Were the tools used on all material by default?
Google	YouTube	CSAI Match	All videos	Upon uploading
	Drive (material that is stored)	CSAI Match	All videos upon uploading and sharing	Upon uploading and viewing
	Drive (material that is shared)	CSAI Match	All videos upon uploading and sharing	Upon sharing and viewing
Meta	Facebook	VMD5 Meta Proprietary technology	All	All material by default
	Facebook Messenger	Meta Proprietary technology	On parts of the service where end-to-end encryption was not enabled (including user reports from end-to-end encrypted threads)	All material by default
	Instagram	VMD5 Meta Proprietary technology VPDQ[23]	Instagram: All parts of the service Instagram Direct: On parts of the service where end-to-end encryption was not enabled (including user reports from end-to-end encrypted threads).	Instagram: All material by default Instagram Direct: All material by default
	Threads	VMD5 Meta Proprietary technology VPDQ	All	All material by default

[23] Meta stated that VPDQ was not used on Instagram Direct.

Microsoft	OneDrive (material when it is shared)	PhotoDNA MD5	For image and video content types that are shared.	All material by default
	Teams	PhotoDNA MD5	All image and video attachments in chat and community threads	All material by default
	Xbox	PhotoDNA MD5	All features where video and imagery can be uploaded or shared.	All material by default
Skype	Skype	PhotoDNA MD5	All videos sent on Skype (excluding Skype Private Conversation)	All material by default
Snap	Spotlight Discover Stories Chat	CSAI Match	Material uploaded from a user's camera roll	Material uploaded from a user's camera roll by default
WhatsApp	WhatsApp (other than Channels)	VideoDNA Internal proprietary technology	Material reported by users	All material by default
	WhatsApp Channels	Internal proprietary technology	All Channel posts	All material by default

Note 1: Apple's iMessage, FaceTime, iCloud, iCloud email; Discord's Discoverable Community Servers, Community Servers, Friend Servers, and Direct Servers; Google's Chat, Google Messages, Gmail; Microsoft's OneDrive (material that is stored) and Outlook.com; and Snap (in Snaps) did not use hash-matching tools to detect known CSEA videos and are therefore not shown in the table. Google stated that Gemini did not allow users to upload or create videos so is not included in table 2.

Note 2: Discord used hash-matching on all direct uploads of GIFs. If a user directly uploaded a GIF to Discord, the same hashing tools to detect known CSEA images processes the first frame of the GIF for matches.

Note 3: Snap stated that in the reporting interval, it used PhotoDNA to proactively scan material uploaded from a user's camera roll to Snapchat by default. This was applied regardless of where in Snapchat a user was seeking to upload an image (for example, whether to Spotlight, Discover, Stories, or Chat). Snap stated that it used this technology to detect known CSEA material on all surfaces of Snapchat where there could reasonably be a hash match (noting that Snaps are novel photos or videos taken using the Snapchat app's camera).

Hash sources

Using the widest pool of reliably verified hash lists gives providers the best chance to detect matches (or near matches), so CSEA material can be removed. This stops the ongoing spread of known CSEA material which could otherwise revicitimise survivors of abuse and harm other members of the public exposed to it.

Many providers that use hash-matching technology, once they detect, verify and hash CSEA material, maintain their own hash lists of that material to prevent it from being re-posted on their services.

There are also expert organisations with extensive hash lists of known CSEA material. These organisations, such as NCMEC, Internet Watch Foundation (**IWF**), Cybertip.ca, as well as law enforcement agencies, make hash lists readily available to industry.

Providers were asked about the sources of hashes they used to detect known CSEA material.

Compared with previous responses[24]

Apple, Discord, Snap, and WhatsApp reported that they used more hash sources since they last received a notice from eSafety.

Discord has greatly improved its hash-matching practices by updating its hash lists daily rather than annually. This allows for recently hashed images to be identified by the service.

Table 3: Hash sources used to detect known CSEA images

Provider	Service	Name all databases	Did the service use all available hashes or a subset from these databases	Frequency the service updated the hashes from these databases
Apple	iCloud email	NCMEC[25] IWF[26] Cybertip.ca[27]	All	Daily
Discord	Discoverable community servers Community servers Friend servers Direct servers	NCMEC	Subset of hashes (vetted hashes)	Daily
		Internal hash database[28]	All hashes	Updated whenever new unknown CSEA material was identified by Discord's human review teams
		IWF hashes that are used for PhotoDNA	All US hashes	Daily
		Threat Exchange hashes for PhotoDNA and PDQ	All hashes	Annually

[24] Non-periodic notices were given in 2022 to Apple, Meta, Microsoft, Skype, Snap and WhatsApp, and in 2023 to Discord and Google. Note that the questions asked and the reporting timeframes differed between notice rounds and between providers. These comparisons serve as a guide to how service responses may have changed since the last notice they received. Future comparisons will be observed through the following three periodic notice responses from providers.
[25] National Center for Missing and Exploited Children
[26] Internet Watch Foundation
[27] Cybertip.ca is the Canadian national tipline for reporting the online sexual exploitation of children
[28] Internal database for all sources of previously detected unknown CSEA images using CLIP

Google	YouTube Drive (material that is stored) Drive (material that is shared) Chat Gemini Gmail	IWF NCMEC Others, including content Google finds on its platforms and subsequently hash.	All hashes Google stated that it reviewed purported CSAM hashes that had not previously been independently reviewed by Google to confirm accuracy. Once Google confirmed it as CSAM, Google inputted it into its detection systems.	External databases: Weekly Internal database: Continuous (every time a hash was matched against content, Google updated the accuracy of the hash)
Meta	Facebook Facebook Messenger Instagram Threads	NCMEC's repository of NGO's CSAM NCMEC's repository of industry's CSAM hashes CSAM hashes from **Lantern** Internal CSEA hash lists generated from Meta's experience reviewing content CSEA hashes from Childline's Report Remove repository	All available hashes	For databases maintained by external parties – At least on an hourly basis For internal databases – Dependent on the frequency with which Meta identified images eligible for banking
Microsoft	OneDrive (material when it is shared) Teams Xbox	Internal database[29] IWF NCMEC	All	Internal database – Every 15 mins External sources - Daily
	Outlook.com	IWF NCMEC	All	Varies[30]
Skype	Skype	Internal database IWF NCMEC	All	Internal database – Every 15 mins External sources - Daily

[29] Internal database of Microsoft-created hashes for CSEA imagery content.
[30] Microsoft stated that the current frequency for updating hash lists on Outlook.com was set at a varying cadence. Microsoft stated that it was improving the frequency at which hash sets were updated and was working to automate this process to ensure it would always using an up-to-date hash set.

Snap	Spotlight, Discover, and Stories	NCMEC's hash list NCMEC Industry hash list NCMEC **Take It Down** IWF **StopNCII** Snap's internally maintained hash list	All available hashes in all databases except those marked as 'Withdrawn/Deleted' in StopNCII	Daily
	Chat	NCMEC's hash list NCMEC Industry hash list NCMEC Take It Down IWF Snap's internally maintained hash list	All available hashes	Daily
WhatsApp	WhatsApp (other than Channels) WhatsApp Channels	NCMEC's NGOs' CSAM hashes NCMEC's industry's CSAM hashes CSAM hashes from separate collaborations within industry Meta's CSEA hash lists WhatsApp's CSAM hash lists CSEA hashes from Childline's Report Remove	All	Internal databases – Frequency depends on the frequency with which the internal source identifies images eligible for banking. External databases – At least hourly

Note: Apple's iMessage, Facetime and iCloud; Google Messages; Microsoft's OneDrive (material that is stored) and Snap's Snaps did not use hash-matching to detect know CSEA and therefore are not included in the table.

Table 4: Hash sources used to detect known child sexual exploitation videos

Provider	Service	Name all databases	Did the service use all available hashes or a subset	Frequency at which the service updated the hashes from these databases
Google	YouTube Drive Meet	NCMEC Others, including content Google finds on its platforms and subsequently hashes	All available hashes are ingested. Google also stated that it reviewed purported CSAM hashes from these databases that had not previously been independently reviewed by Google to confirm accuracy. Once Google confirmed it as CSAM, Google inputted it into its detection systems.	External databases: One ingestion from NCMEC during the report period Internal database: Continuous (every time a hash was matched against content, Google updated the accuracy of the hash)
Meta	Facebook Facebook Messenger Instagram Threads	NCMEC's repository of NGO's CSAM NCMEC's repository of industry's CSAM hashes CSAM hashes from Lantern Internal CSEA hash lists generated from Meta's experience reviewing content CSEA hashes from Childline's Report Remove repository CSEA hashes from the Korean Communication Standards Commission hash list that was not yet incorporated into the NCMEC NGO list	All available hashes	For databases maintained by external parties – At least on an hourly basis For internal databases – Dependent on the frequency with which Meta identified images eligible for banking
Microsoft	OneDrive (material when it is shared) Teams Xbox	Internal database IWF NCMEC	All	Internal – Every 15 mins External – Daily

Skype	Skype	Internal database IWF NCMEC	All	Internal – Every 15 mins External – Daily
Snap	Spotlight Discover Stories Chat	CSAI Match	All	Continuously (Snap calls the CSAI Match database upon scanning material uploaded from a user's camera roll, so it uses the current hash list from that database at all times.)
WhatsApp	WhatsApp (other than Channels) WhatsApp Channels	NCMEC's NGOs' CSAM hashes NCMEC's industry's CSAM hashes CSAM hashes from separate collaborations within industry Meta's CSEA hash lists WhatsApp's CSAM hash lists CSEA hashes from Childline's Report Remove	All	Internal databases – Frequency depends on the frequency with which the internal source identifies images eligible for banking. External databases – At least hourly

Note 1: Apple's iMessage, FaceTime, iCloud and iCloud email; Discord's Discoverable Community Servers, Community Servers, Friend Servers, Direct Servers; Google's Chat, Google Messages and Gmail; Microsoft's OneDrive (material that is stored) and Outlook.com; and Snap's Snaps did not use hash-matching tools to detect known CSEA videos. Google stated that Gemini did not allow users to upload or create videos so is not included in table 4.

Note 2: Snap stated that Snaps are novel photos or videos taken using the Snapchat app's camera, which can be distinguished from images and videos that are uploaded from a user's device camera roll.

Note 3: Discord used hash-matching on all direct uploads of GIFs. If a user directly uploaded a GIF to Discord, the same hashing tools used to detect known CSEA images processed the first frame of the GIF for matches. See table 3 for a list of databases Discord is using to detect known CSEA images

Blocking URLs to known CSEA material

Blocking **URLs** to known CSEA material is an important safety measure in preventing users from accessing and spreading CSEA. eSafety investigators are aware of services being used to distribute thousands of URLs to sites that host CSEA material. Some URLs are posted under the guise of providing access to lawful adult pornography however some URLs are posted and advertised explicitly as providing access to CSEA material.

Not-for-profit organisations such as the Internet Watch Foundation (IWF) have **URL databases** (lists of web addresses) of URLs linking to previously reported CSEA images and videos, which

they make available to providers.[31] Law enforcement agencies around the world also have databases of URLs to known CSEA material. Blocking URLs such as these is a common practice across many services to protect user safety and security, as well as to comply with various laws.

Despite the availability of databases that identify URLs which link to known CSEA material and websites that are dedicated to hosting it, some providers are not using them.

Providers were asked if their services were blocking URLs to known CSEA material and, if so, what the sources of the URLs were.

Key insights:

- Apple, Discord, Google and WhatsApp did not block URLs to known CSEA material on any part of their services.
- Discord did not block URLs to known CSEA material on any part of its services, except for messages reported to Discord, despite joining the IWF in 2023 and informing eSafety that this membership was to facilitate Discord in starting to block URLs to known CSEA.
- Microsoft did block URLs to known CSEA material on Xbox but not on Outlook or Teams.

Compared with previous responses[32]

eSafety previously reported on the inaction of Discord, Google and WhatsApp in blocking URLs to known CSEA material on their services in our 2022 and 2023 transparency reports on CSEA.

[31] Internet Watch Foundation, 2023, 'URL List', accessed 16 July 2025, URL: https://www.iwf.org.uk/our-technology/our-services/url-list/

[32] Non-periodic notices were given in 2022 to Apple, Meta, Microsoft, Skype, Snap and WhatsApp, and in 2023 to Discord and Google. Note that the questions asked and the reporting timeframes differed between notice rounds and between providers. These comparisons serve as a guide to how service responses may have changed since the last notice they received. Future comparisons will be observed through the following three periodic notice responses from providers.

Table 5: Of the providers who were blocking URLs from known CSEA material, what were the sources of these URLs?

Provider	Service	Source of these URLs
Meta	Facebook Facebook Messenger Instagram Threads	Lantern IWF URLs from material which violated its CSEA policies, that Meta's integrity team identified were added to its internal database
Microsoft	Xbox	Lantern The Tech Coalition GIFCT NCMEC Open-sourced information
Snap	Spotlight Discover Chat Stories Snaps	Lantern URLs that Snap internally identified with a high degree of confidence linked to CSEA material

Note: Apple iMessage, FaceTime, iCloud, iCloud email; Discord's Discoverable Community Servers, Community Servers, Friend Servers, Direct Messages; Google's YouTube, Drive, Meet, Chat, Messages and Gmail; Microsoft's Outlook.com and Teams; Skype; WhatsApp (other than Channels) and WhatsApp Channels are not included in table 5 as they did not block URLs linking to known CSEA material.

Proactively detecting new CSEA material

Hash-matching tools can only 'match' against previously identified and confirmed ('known') CSEA material and seek to prevent its ongoing dissemination. Other tools can be deployed on services to prevent the sharing of new CSEA, and before it has been verified and included in a database. These tools help stop the abuse of children and the spread of the CSEA, and alert providers to users who are sharing it or coercing children to produce it, which is illegal.

Tools to detect new CSEA may use artificial intelligence (AI) 'classifiers' to identify material that is likely to depict the abuse of a child, and typically to prioritise these cases for human review and verification. These tools are trained on various datasets, including verified CSEA, as well as material that does not contain CSEA, in order to identify the markers of content depicting abuse. An example of this technology is Google's Content Safety API[31] or Thorn's classifier, which Thorn reports has a 99% precision rate.[32]

The deployment of tools is important across the various online spaces in which new CSEA material may be generated, stored and/or disseminated. For example, offenders may look to coerce children into producing and sharing CSEA in private online spaces, such as direct messaging functions. In 2022, Thorn reported that 2 in 3 children had been asked by someone

online to move from a public chat into a private conversation on a different platform.[33] A 2023 report by Thorn also found that Snapchat was a common platform for sending self-generated CSAM, perhaps due to features such as 'disappearing' direct messages. 39% of 13 to 17 year olds who had shared their own nudes did so via direct messages 'in apps where content disappears, like Snapchat'.[34]

INHOPE reported that the reports it received of new records in 2024 (929,733) reflected a 34.8% increase compared to 2023. It stated that this 'surpassed the peak levels seen in 2021 (760,054) by 22.4%, demonstrating a significant rise in the identification of previously unseen materials'[35]

Providers were asked about the tools they were using to detect new CSEA for both images and videos.

Key insights:

- Apple did not use tools to detect new CSEA on iCloud (which was end-to-end encrypted based on whether a user opted in) and iCloud email. For iMessage and FaceTime (which are end-to-end encrypted) Apple only used Communication Safety, Apple's safety intervention to identify images or videos that likely contain nudity, as a means of 'detecting' CSEA.
- Google did not use tools to detect new CSEA on Google Meet, Google Chat, Google Messages (which was end-to-end encrypted) or Gmail. Although Google used **machine learning classifiers** to detect new CSEA material on Google Drive, Google only did this after an account was flagged as 'suspicious'.
- Microsoft did not use tools to detect new CSEA images or videos on OneDrive, Outlook or Teams.
- Skype did not use tools to detect new CSEA material.
- Although Snap used Google's Content Safety API and **machine learning** nudity detection to detect new CSEA images and videos, this usage was limited to material reported to Snap.[36] Tools were not used to detect new CSEA in material on Private Stories and Chat (which are private surfaces), unless it was reported to Snap.

---

[33] Thorn, 2022, 'Online Grooming: Examining risk encounters amid everyday digital socialization', accessed 17 July 2025, URL: https://info.thorn.org/hubfs/Research/2022_Online_Grooming_Report.pdf?_gl=1*118xry7*_gcl_au*MTIzNDY3NzQ2OS4xNzUyNzM0NDc5

[34] Thorn, 2023, LGBTQ+ Youth Perspectives: How LGBTQ+ Youth are Navigating Exploration and Risks of Sexual Exploitation Online', accessed 17 July 2025, https://www.thorn.org/research/library/lgbtq-teens-are-at-a-greater-risk-for-exploitation-online/, Thorn, 'Trends in Financial Sextortion', 2024, accessed 13 June 2025, URL: https://www.thorn.org/research/library/financial-sextortion/

[35] INHOPE, 2025, 'Annual Report 2024' accessed 11 June 2025, URL: https://inhope.org/EN/articles/annual-reports

[36] Google's Content Safety API was only used on material (including Snaps) on Spotlight, Discover, and Public Stories, and media (including Snaps) reported to Snap, in Private Stories and Chat reported to Snap.

Compared with previous responses[37]

eSafety previously reported on the lack of tools used for detecting new CSEA on the services mentioned in the Key Insights in our 2022 and 2023 transparency reports on CSEA.

While gaps remain, eSafety observed positive changes on some services:

In 2022, Snap reported that it did not use any tools to detect new CSEA. It is therefore a positive improvement that Snap now reported using tools to detect new CSEA.

In 2022, Microsoft reported that it did not use any tools to detect new CSEA but it was using classifiers on OneDrive, Teams and Xbox Live to detect nudity, which may also identify some CSEA content. It is therefore a positive improvement that Microsoft now reported using tools to detect new CSEA on Xbox. Microsoft could continue to improve its practice by expanding its use of tools to its other services.

Discord and Meta reported using more tools to detect new CSEA than they did in their previous responses to the notices given to them in 2022 and 2023 respectively.

Table 6: Tools providers were using to detect new CSEA

Provider	Service	Name all tools used to detect new CSEA images	Name all tools used to detect new CSEA videos	Were the tools used on all or only parts of the service?	Were the tools used on all material by default?
Apple	iMessage	Communication Safety	Communication Safety	All	To detect nudity in images and videos
	FaceTime	Communication Safety	Communication Safety	FaceTime video messages	

[37] Non-periodic notices were given in 2022 to Apple, Meta, Microsoft, Skype, Snap and WhatsApp, and in 2023 to Discord and Google. Note that the questions asked and the reporting timeframes differed between notice rounds and between providers. These comparisons serve as a guide to how service responses may have changed since the last notice they received. Future comparisons will be observed through the following three periodic notice responses from providers.

Discord	Discoverable community servers Community servers Friend servers Direct servers	CLIP	Google's Content Safety API	All messages sent in Discord servers and via direct messages, banners, server avatars, and user avatars	By default
Google	YouTube	Machine learning classifiers (Google's Content Safety API)	Machine learning classifiers (Google's Content Safety API)	All parts of the service	By default
	Drive (material that is stored) Drive (material that is shared)	Machine learning classifiers (Google's Content Safety API)	Machine learning classifiers (Google's Content Safety API)	For Consumer only	After an account had been flagged as suspicious
	Gemini	Machine learning classifiers (Google's Content Safety API)	N/A	All images	By default
Meta	Facebook	CSAM content classifier CSAM solicitation content classifier Child sexualisation content classifier FB group classifier	CSAM content classifier CSAM solicitation content classifier Child sexualisation content classifier FB group classifier	All	All material by default
	Facebook Messenger	Strikable multi-class classifier Account-disable multi-class classifier Child grooming binary classifier	Strikable multi-class classifier Account-disable multi-class classifier Child grooming binary classifier	User reports	

	Instagram Threads	Instagram, Threads: CSAM content classifier CSAM solicitation content classifier Child sexualisation content classifier Instagram Direct: Strikable multi-class classifier Account-disable multi-class classifier Child grooming binary classifier	Instagram, Threads: CSAM content classifier CSAM solicitation content classifier Child sexualisation content classifier Instagram Direct: Strikable multi-class classifier Account-disable multi-class classifier Child grooming binary classifier	Instagram: All Threads: All Instagram Direct: user reports	
Microsoft	Xbox	WatchFor[38]	WatchFor	All imagery	All material by default
Snap	Spotlight	Content Safety API Machine Learning Nudity Detection	Content Safety API Machine Learning Nudity Detection	Proactively scanned material uploaded to Spotlight, including material on Spotlight that was reported to Snap	All material by default and on material reported to Snap
	Discover	Content Safety API Machine Learning Nudity Detection	Content Safety API Machine Learning Nudity Detection	Proactively scanned material uploaded to Discover, including material on Discover that was reported to Snap	All material by default and on material reported to Snap
	Stories	Content Safety API Machine Learning	Content Safety API Machine Learning	Proactively scanned material	All material posted on Public

[38] Microsoft internal tool based on AI models used for processing and analysing images, videos, live streams, GIFs, audio and text.

		Nudity Detection	Nudity Detection	uploaded to Public Stories, including material on Public and Private Stories that was reported to Snap	Stories by default and on material reported to Snap
	Chat	Content Safety API Machine Learning Nudity Detection	Content Safety API Machine Learning Nudity Detection	Chat material that was reported to Snap	Chat material reported to Snap
	Snaps	Content Safety API Machine Learning Nudity Detection	Content Safety API Machine Learning Nudity Detection	Proactively scanned Snaps posted on public surfaces (for example, Public Stories, Map Stories, Discover and Spotlight), as well as Snaps reported to Snap.	Snaps posted on public surfaces (for example, Public Stories, Map Stories, Discover, Spotlight) and on Snaps reported to Snap.
WhatsApp	WhatsApp (other than Channels)	WhatsApp's CSAM Image Classifier Google's Content Safety API	WhatsApp's CSAM Image Classifier Google's Content Safety API	Material reported by users Groups profile picture Communities profile picture	All material by default
	WhatsApp Channels	WhatsApp's CSAM Image Classifier Google's Content Safety API	WhatsApp's CSAM Image Classifier Google's Content Safety API	Channel profile picture Channel posts	

Note 1: Apple's iCloud and iCloud email; Google's Meet, Chat, Messages and Gmail; Microsoft's OneDrive (material that it is stored), OneDrive (material when it is shared), Outlook.com and Teams; and Skype did not use tools to detect new CSEA in images or videos and therefore are not included in the table.

Note 2: Google's Gemini did not allow users to upload or create videos during the report period and is therefore not included in the table.

Detecting CSEA livestreaming

CSEA livestreaming is the transmission or receipt of live acts of sexual exploitation or abuse of children via webcam or video to people anywhere in the world, whether or not in exchange for payment. CSEA livestreaming includes one-on-one video calls and video calls where one or more people stream CSEA material and/or activity to a group of any size. This can be through a video call between two or more participants, or through a livestream that is broadcast to multiple users. This can occur where a sexual offender controlling the child charges money for providing access to the livestreaming[39] or when a child is coerced into performing live sexual acts by online offenders, who often record and share the videos elsewhere.[40]

Some providers that were asked about steps to detect **CSEA livestreaming** drew a distinction between 'livestreaming' and 'video calls' or 'video conferences'. eSafety is aware of live CSEA activity taking place in both one-to-one video calls as well as on livestreaming features that allow multiple participants. Research of livestreamed abuse of children in the Philippines found that it typically took place in a secure environment with an encrypted connection between two or more parties.[41]

Research and analysis by the Australian Institute of Criminology (AIC) identified online **video calling services** as a central vector for livestreaming CSEA.[42] The International Justice Mission (IJM) and University of Nottingham Rights Lab found that video chat and messaging services commonly used to facilitate livestreamed child sexual abuse are either optionally end-to-end encrypted, end-to-end encrypted by default, or moving towards end-to-end encryption.[43] Detecting CSEA activity in a live video is more technically challenging than detecting CSEA material in still images, given the volume of content transmitted. However, there are tools available to detect when CSEA is occurring, and **language analysis tools** that work by assessing the words and phrases used while a video is being streamed, or by using signals outside of a call to detect when CSEA activity is likely to be in a live video (this is useful for end-to-end encrypted services).[44]

---

[39] Australian Institute of Criminology, 2023, 'The overlap between child sexual abuse live streaming, contact abuse and other forms of child exploitation', accessed 26 May 2025, URL: https://www.aic.gov.au/publications/tandi/tandi671

[40] Australian Federal Police, 'AFP warn about fast growing online child abuse trend', published 11 September 2021, accessed 25 March 2025, URL: https://www.afp.gov.au/news-centre/media-release/afp-warn-about-fast-growing-online-child-abuse-trend

[41] WeProtect Global Alliance, 2025, 'Global Threat Assessment 2023 Assessing the scale and scope of child sexual abuse online, accessed 28 May 2025, URL: https://www.weprotect.org/wp-content/uploads/Global-Threat-Assessment-2023-English.pdf

[42] Australian Institute of Criminology, 2021, 'Submission to Inquiry into Law Enforcement Capabilities in Relation to Child Exploitation', accessed 22 May 2025, URL: https://www.aph.gov.au/DocumentStore.ashx?id=5b62f90b-b0e2-42ef-ab27-0225cff04a29&subId=711815

[43] WeProtect Global Alliance, 2025, 'Global Threat Assessment 2023 Assessing the scale and scope of child sexual abuse online, accessed 28 May 2025, URL: https://www.weprotect.org/wp-content/uploads/Global-Threat-Assessment-2023-English.pdf

[44] eSafety is aware of AI and machine learning technologies that work to detect and block CSEA on video livestreams on-device, without compromising user privacy, and some with end-to-end encryption applications.

Beyond detection technologies, providers can also put in place measures such as safety prompts or age assurance measures to discourage or prevent unsafe interactions between child users and adult users.

Despite the availability of technology to help detect child sexual exploitation and abuse in livestreams or video calls, not all providers are using it.

Providers whose services offer video call or livestreaming capabilities were asked if they used tools to detect CSEA livestreaming.

Key insights:

- Apple did not use any tools to detect CSEA livestreaming on FaceTime.
- Discord did not use any tools to detect CSEA livestreaming on Go Live or Video Calls (voice and video calls became end-to-end encrypted in September 2024).
- Google did not use tools to detect CSEA livestreaming on Google Meet, but did use tools on YouTube.
- Meta did not use tools to detect CSEA livestreaming on Facebook Messenger, but did use tools on Facebook live.
- Microsoft did not use tools to detect CSEA livestreaming Teams.
- Skype did not use tools to detect CSEA livestreaming.
- Snap did not use tools to detect CSEA livestreaming in Snapchat Video Chats.
- WhatsApp did not use tools to detect CSEA livestreaming.

Compared with previous responses[45]

eSafety previously reported on the lack of tools used for detecting CSEA livestreaming on Apple's FaceTime, Discord's livestreams and voice chats, Microsoft Teams, and Skype in our 2022 and 2023 transparency reports on CSEA.

[45] Non-periodic notices were given in 2022 to Apple, Meta, Microsoft, Skype, Snap and WhatsApp, and in 2023 to Discord and Google. Note that the questions asked and the reporting timeframes differed between notice rounds and between providers. These comparisons serve as a guide to how service responses may have changed since the last notice they received. Future comparisons will be observed through the following three periodic notice responses from providers.

Table 7: Tools used to detect CSEA livestreaming

Provider	Service	Names of all tools used to detect the livestreaming of CSEA	Were the tools used on all livestreams by default, or only in certain circumstances?
Google	YouTube	A combination of machine learning language analysis, video classifiers, and human moderators	All material by default
Meta	Facebook Live[46]	Media Match Service Internal proprietary classifier	All material by default
	Instagram Live[47]	Media Match Service Meta proprietary tool	

Note 1: Apple's FaceTime; Discord's Go Live and Video calls; Google Meet; Meta's Facebook Messenger video calls; Microsoft's Teams; Skype, Snap's Snapchat Video Chats; and WhatsApp's video calls did not use tools to detect CSEA livestreaming and are therefore not shown in the table.

Note 2: Microsoft's Xbox did not provide video call or livestreaming functionality and is therefore not shown in the table.

Note 3: Please see WhatsApp's provider section to see additional information on WhatsApp's detection of CSEA livestreaming.

Providers were then asked to provide details of the languages the tools operated in. CSEA livestreaming identified through machine learning tools typically requires review by human moderators to verify abusive material and activity flagged by technology or reported by users. Identifying some harms (such as CSEA livestreaming) requires review of the surrounding conversation, as well as behavioural signals, to confirm that CSEA activity is taking place.

The Australian Bureau of Statistics 2021 Census results list Mandarin, Arabic, Vietnamese, Cantonese and Punjabi as the top languages spoken in Australian homes, other than English.[48] Not all providers operate tools across these five languages, let alone other major languages spoken in Australia and by the providers' wider users. This is important as harmful activities can go undetected if moderators cannot operate in the languages spoken by Australians.

[46] Meta stated that once the livestream was complete, the videos were marked as 'was live' and stored as videos on demand. The measures identified in tables 4 and 6 then applied to these videos.
[47] Meta stated that once the livestream was complete, the videos were marked as 'was live' and stored as videos on demand. The measures identified in tables 4 and 6 then applied to these videos.
[48] Australian Bureau of Statistics, 2022, 'Cultural diversity: Census', accessed 11 April 2025, URL: https://www.abs.gov.au/statistics/people/people-and-communities/cultural-diversity-census/latest-release#:~:text=Key%20statistics,-27.6%20per%20cent&text=61%2C860%20international%20visitors%20were%20in,Punjabi%20(0.9%20per%20cent).

Many tools to identify harms like grooming only operate in English, or a small number of languages. Where this is the case, it is particularly important that providers have human moderators operating in the languages of the communities they offer services to. Relying on language translation tools risks losing important context and cultural or linguistic nuance.

Table 8 provides a summary of the languages the tools to detect CSEA livestreaming operated in. For a full list of all languages covered, please see the individual provider responses at the end of the report. The languages which are spoken by providers' human moderators can be found in the 'Additional information from individual providers' section from page 94.

Table 8: Languages covered by tools used to detect livestreaming of CSEA activity

Provider	Service or part of the service	Languages the tools operated in
Google	YouTube	104 languages
Meta	Facebook Live Instagram Live	Audio transcription: 22 languages Optical character recognition (OCR): 39 languages Text analysis through an internal proprietary tool: 99 languages

Note 1: Apple's FaceTime, Discord's Go Live and Video calls, Google's Meet, Meta's Facebook Messenger Video Calls, Microsoft's Teams; Skype; Snap's Snapchat Video Chats; and WhatsApp video calls did not use tools to detect CSEA livestreaming and therefore are not shown in the table.

Note 2: Microsoft's Xbox did not provide video call or livestreaming functionality and is therefore not shown in the table.

Proactively detecting CSEA activity – grooming

eSafety is aware that offenders use services to 'groom' children. Grooming is predatory conduct to prepare a child for sexual activity, so they can be coerced into meeting the offender, or to send images or videos to the offender, or engage in sexual activity through video livestreams. Grooming is a form of child sexual abuse and exploitation.

The WeProtect Global Alliance reported that online grooming, is a key concern.[49] Research suggests that it takes just 45 minutes on average for a child to be groomed in a social gaming environment, with the shortest recorded time being 19 seconds in just 7 messages.[50] Perpetrators attempting to groom children online may contact multiple children simultaneously, knowing a small percentage will respond. They use social media, chat rooms, gaming environments and other platforms with messaging functions to identify possible targets. They

[49] WeProtect Global Alliance, 2025, 'Global Threat Assessment 2023 Assessing the scale and scope of child sexual abuse online, accessed 28 May 2025, URL: https://www.weprotect.org/wp-content/uploads/Global-Threat-Assessment-2023-English.pdf

[50] WeProtect Global Alliance, 2025, 'Global Threat Assessment 2023 Assessing the scale and scope of child sexual abuse online, accessed 28 May 2025, URL: https://www.weprotect.org/wp-content/uploads/Global-Threat-Assessment-2023-English.pdf

may then divert conversations to an end-to-end encrypted environment to lower their risk of detection, as fewer tools to detect this harm can operate in these environments. This is known as 'off platforming'. In 2025[51] and 2024, the Australian Federal Police reported convictions of child abuse related to grooming on online services.[52]

Providers were asked if they used tools to detect grooming on their services.

Key insights:

- Apple did not use language analysis tools to detect grooming in iCloud email, iMessage or FaceTime.
- Discord did use language analysis tools to detect grooming of children in direct messages (but not on discoverable community servers, community servers or friend servers).
- Google did not use language analysis tools to detect grooming of children on Meet, Chat or Messages.
- Microsoft did use language analysis tools to detect grooming on Xbox, but not on Teams.
- Skype did not use language analysis tools to detect grooming.
- WhatsApp did not use language analysis tools to detect grooming on WhatsApp messages (which was end-to-end encrypted), but did use tools to detect grooming in Channels.
- Snap used language analysis tools to detect grooming but only on material and accounts reported to Snap (not proactively).

[51] Australian Federal Police, 2025, ' Victorian man jailed for online grooming and child exploitation offences', accessed 13 March 2025, URL: https://www.afp.gov.au/news-centre/media-release/victorian-man-jailed-online-grooming-and-child-exploitation-offences

[52] Australian Federal Police, 2024, 'Man jailed for ongoing sexual abuse and grooming of overseas-based child', accessed 13 March 2025, URL: https://www.afp.gov.au/news-centre/media-release/man-jailed-ongoing-sexual-abuse-and-grooming-overseas-based-child

Compared with previous responses[53]

eSafety previously reported that the services mentioned in the Key Insights did not use language analysis tools for detecting grooming on in our 2022 and 2023 transparency reports on CSEA.

While gaps remain, eSafety observed positive changes on some services:

In 2022, Snap reported that it did not use any tools to detect grooming. It is therefore a positive improvement that Snap now reported using tools to detect grooming. Snap could continue to improve its practice by using tools to detect grooming proactively on all relevant parts of its service, rather than only on content reported by users.

In 2023, Discord reported that it did not use any tools to detect grooming. It is therefore a positive improvement that Discord now reported using tools to detect grooming of children in direct messages. Discord could continue to improve its practice by using tools to detect grooming on more parts of its service.

Additionally, Meta reported that it now used more language analysis tools to detect grooming on Facebook Messenger and Instagram Direct than it did in 2022.

Use of language analysis technology to detect grooming

As with CSEA livestreaming, language analysis is an important tool to employ in proactively detecting terms, abbreviations, codes and hashtags that indicate likely grooming. Language analysis tools can assign a probability to written messages (such as texts) allowing them to be flagged for human review.[54] Harms such as grooming, are highly context dependent and require review of the surrounding conversation, as well as behavioural signals, to confirm the CSEA activity is taking place.

Without automated language analysis capabilities, there is a risk that grooming may continue and remain unaddressed if the child targeted does not recognise and report the predatory contact themselves.

Table 9 shows the providers that were using language analysis technologies to detect grooming and the technologies they were using.

---

[53] Non-periodic notices were given in 2022 to Apple, Meta, Microsoft, Skype, Snap and WhatsApp, and in 2023 to Discord and Google. Note that the questions asked and the reporting timeframes differed between notice rounds and between providers. These comparisons serve as a guide to how service responses may have changed since the last notice they received. Future comparisons will be observed through the following three periodic notice responses from providers.

[54] Child Rights International Network, 2023, 'Explaining the technology for detecting child sexual abuse online', accessed 5 March 2025, URL: https://home.crin.org/readlistenwatch/stories/explainer-detection-technologies-child-sexual-abuse-online

Table 9: Language analysis technologies used to detect likely terms, abbreviations, codes and hashtags indicating grooming

Provider	Service	Name of language analysis technologies
Discord	Direct messages	Safety Alerts in Chat Sensitive Media Filter[55]
Google	YouTube	Machine learning classifiers
Meta	Facebook Facebook Messenger	Internal proprietary tools
	Instagram Threads	Instagram, Threads: Internal proprietary tools Instagram Direct: Internal proprietary tools
Microsoft	Xbox	Community Sift[56] Internal proprietary tool B Internal proprietary tool C
Snap	Spotlight Discover Stories Chat Snaps	Signals-based detection Abusive language detection Keyword-based detection
WhatsApp	WhatsApp Channels	Internal proprietary tool A Internal proprietary tool B

Note: Apple's iCloud email, FaceTime and iMessage; Discord's Discoverable Community Servers, Community Servers and Friend Servers; Google's Meet, Chat and Messages; Microsoft Teams; Skype; and WhatsApp (other than Channels) are not included in table 9 as they did not use language analysis technology to detect grooming.

Languages analysed

To be able to detect grooming, it's important that service providers have language analysis tools that operate in the languages of the communities that use their services. The most common languages spoken in Australian homes, other than English, are Mandarin, Arabic, Vietnamese, Cantonese and Punjabi.[57]

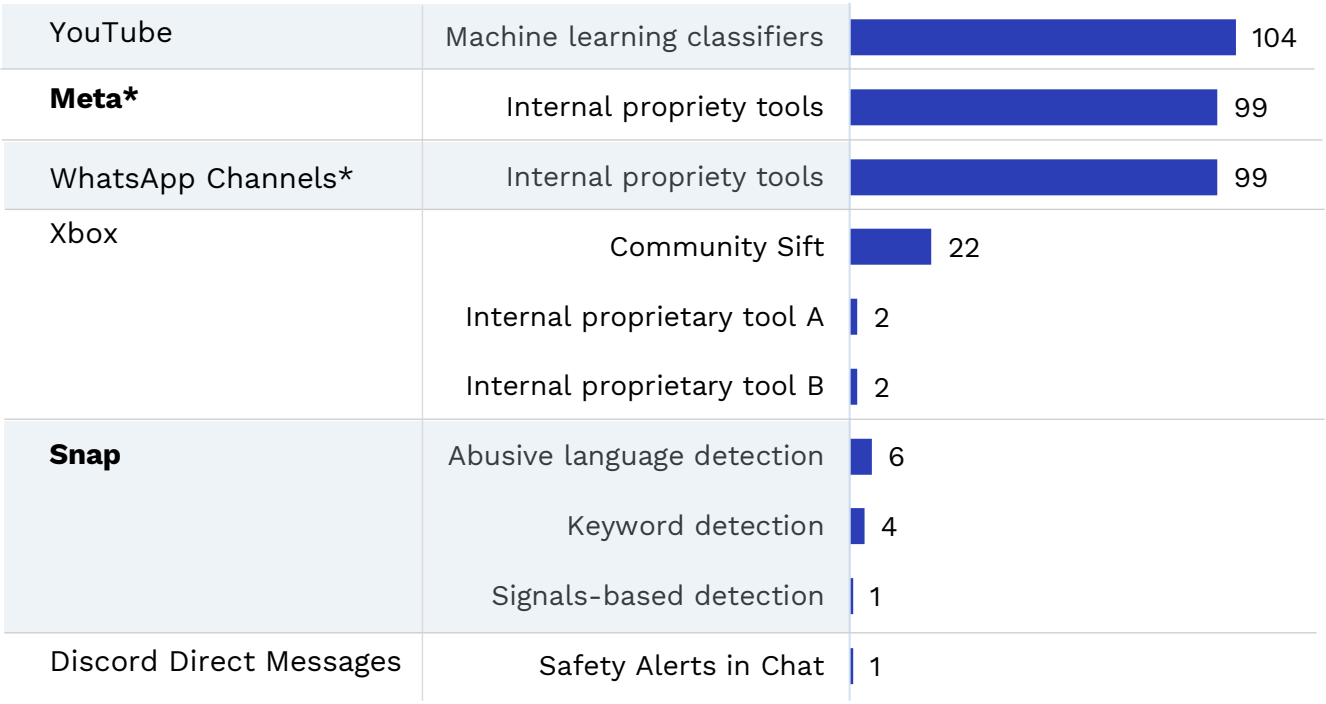
Figure 2 provides a summary of the languages the tools operated in. For a full list of all languages please see the individual provider responses at the end of the report.

[55] Safety Media Filter is a filter tool which automatically blurs potentially sensitive media sent to teens in direct messages and in servers. The filter processes image-based media posted on Discord and blurs or blocks content when it detect that the content might contain nudity, graphic sex acts, or sexually explicit material.
[56] Community Sift is an AI-powered content moderation platform, see: https://developer.microsoft.com/en-us/games/products/community-sift/#accordion-46019bb432-item-6a6aac48a9
[57] Australian Bureau of Statistics, 2022, 'Cultural diversity: Census', accessed 11 April 2025, URL: https://www.abs.gov.au/statistics/people/people-and-communities/cultural-diversity-census/latest-release#:~:text=Key%20statistics,-27.6%20per%20cent&text=61%2C860%20international%20visitors%20were%20in,Punjabi%20(0.9%20per%20cent).

Key insights:

- Google's language analysis tools to detect grooming for YouTube operated in 104 languages, however, the list of languages included the classical language Latin, alongside two invented languages (Volapük and Ido)[58] but did not include Cantonese[59] (see page 116 for the full list).
- Discord's language analysis tools to detect grooming only operated in English.
- Microsoft used multiple language analysis tools to detect grooming on Xbox, but two of these tools only covered two languages. The third tool covered 22 languages but did not include Punjabi or Cantonese (see page 134 for the full list).
- Snap's language analysis tools to detect grooming did not operate in Mandarin, Vietnamese, Cantonese or Punjabi.

Figure 2: Number of languages covered by analysis technologies used to detect likely terms, abbreviations, codes and hashtags indicating grooming



*Meta's Facebook, Facebook Messenger, Instagram and Threads and WhatsApp Channels utilised an additional internal proprietary language agnostic tool. The languages covered by the tool depended on the languages of the words or phrases entered into the tool.

Note 1: Apple's iCloud email, FaceTime and iMessage; Discord's Discoverable Community Servers, Community Servers and Friend Servers; Google's Meet, Chat and Messages; Microsoft Teams; Skype; and WhatsApp (other than Channels) are not included in Figure 2 as they did not use language analysis technology to detect grooming.

Note 2: Meta includes Facebook, Facebook Messenger, Instagram and Threads.

[58] Invented (or constructed) languages are languages that are designed and created artificially, rather than evolving naturally through use. Invented languages are not typically spoken as native languages.
[59] Google's tools covered Chinese (Traditional) but not specifically Cantonese.

Note 3: Snap includes Stories, Spotlight, Discover, Chat and Snaps

Note 4: A list of the languages can be found in the 'Additional questions to providers' section for each of the providers.

Providers are able to use a number of sources to aid with detecting terms, abbreviations, codes and hashtags that indicate grooming. Organisations specialising in online child safety such as NCMEC, Thorn and Lantern make lists of these for providers to use on their services.

Providers were asked about the sources they used to aid with the detection of grooming.

Table 10: Organisations used to source terms, abbreviations, codes and hashtags indicating grooming

Provider	All the organisations the service(s) used to source terms, abbreviations, codes and hashtags
Discord	Thorn
Meta	Lantern, and Meta's own ongoing integrity work
Microsoft	Xbox: Global Internet Forum to Counter Terrorism (GIFCT); National Center for Missing and Exploited Children (NCMEC), Tech Coalition, Internet Watch Foundation and Lantern
Snap	Lantern
WhatsApp	Meta's own ongoing integrity work.

Note 1: Apple, Google, Microsoft Teams and Skype were not utilising external organisations to source terms, abbreviations, codes and hashtags indicating grooming and therefore not shown the in the table.

Note 2: Google stated that it did not source specific terms or abbreviations from external organisations, or simply match words to a known list of violative terms, as language in this context evolves quickly.

Steps taken if grooming was flagged

While language analysis tools are a valuable resource in identifying potential grooming, review by human moderators may be required to verify content flagged by those tools. This is particularly important where the terms, abbreviations, codes and hashtags have not been identified previously or where effective moderation depends on understanding the language and culture for context – as is the case for harms such as grooming. It is therefore particularly important that companies have human moderators operating in the languages of the communities they offer services to. Relying on language translation tools risks losing important context and cultural or linguistic nuance.

Providers were asked if human moderators reviewed all reports of potential grooming flagged by the language analysis tools, if the tools were used on all material by default, and what steps were taken on the account when the language analysis technology detected grooming.

Consistent with other providers, Discord's notice required it to report the proportion of reports flagged by language analysis technology used to detect grooming that were reviewed by human moderators. Discord stated that its human moderators reviewed a statistically significant sample of the communications identified by its safety measure, Safety Alerts in Chat, to assess its effectiveness but did not state the proportion of these reports flagged that were reviewed by human moderators.

Consistent with other providers, Google's notice required it to report the proportion of reports flagged by language analysis technology used to detect grooming that were reviewed by human moderators. Google stated that human moderators reviewed a subset of reports flagged by its language analysis technology to detect grooming on YouTube, but that the proportion of reports reviewed by human moderators was not available because YouTube's flags for human moderator review were not specific to the indicators flagged by language analysis technology to detect grooming.

Consistent with other providers, WhatsApp's notice required it to report the proportion of reports flagged by language analysis technology used to detect grooming that were reviewed by human moderators. WhatsApp stated that its language analysis tools did not generate a binary flag as to whether a report was or was not likely to involve inappropriate interactions with a child (or any other violation type). Instead, these tools generated a score which reflected the likelihood that a report may violate one or more of WhatsApp's policies, which was then combined with signals (including signals from language analysis tools as well as others) to create an overall prioritisation score. The overall prioritisation score enabled WhatsApp to prioritise reports relating to high severity violations. There was no threshold or cut off score that automatically resulted in human review. The system works in a dynamic manner with all reports being enqueued in order of priority, based on the overall prioritisation score, and the order adjusted as new reports enqueued. WhatsApp was unable to isolate reports where language analysis tools contributed to the overall prioritisation score.

Table 11: How the language analysis technology was used

Provider	Service	Was the language analysis technology used on all material by default, or only in certain circumstances?	What steps did the service take when the language analysis technology detected grooming on an account?
Services that had a human moderator review all reports flagged by the language analysis technology			
Meta[60]	Facebook Threads	Internal proprietary tool A: All material by default. Internal proprietary tool B: Only used to help facilitate human review of user reports.	For classifiers using internal proprietary tool A: Depending on signals and confidence of the classifier, content was either automatically deleted or enqueued for human review. Meta may have also taken enforcement action at the account level. For classifiers using internal proprietary tool B: If the tool detected a word or phrase associated with grooming, it highlighted the word or phrase to facilitate human review of the relevant material. If the human reviewer then confirmed a violation of Meta's policies, took appropriate enforcement action against the material and/or account.
	Facebook Messenger	Internal proprietary tool A: Only on Meta's public messaging products (Channels and Community Chats) and on other material reported by users.[61] Internal proprietary tool B: Only used to help facilitate human review of user reports.	
	Instagram	Instagram: Internal proprietary tool A: All material by default. Internal proprietary tool B: Only used to help facilitate human review of user reports. Instagram Direct: Internal proprietary tool A: Only on Meta's public messaging products (Channels and Community Chats) and on other material reported by users. Internal proprietary tool B: Only used to help facilitate human review of user reports.	

[60] Meta stated that language analysis tools did not generate a binary flag as to whether a report was or was not likely to be inappropriate interactions with a child or any other violation type. Instead, these tools generate a score which reflects the likelihood that a report may violate one or more of Meta's policies, which is then combined with signals (including signals from language analysis tools as well as others) to create an overall prioritisation score. The overall prioritisation score enabled Meta to prioritise reports relating to high severity violations. There was no threshold or cut off score that automatically resulted in human review. The system worked in a dynamic manner with all reports being enqueued in order of priority, based on the overall prioritisation score, and the order adjusted as new reports enqueued. Meta was unable to isolate reports where language analysis tools contributed to the overall prioritisation score.

[61] Meta stated that different classifiers leveraging this tool may have covered different surfaces (for example Meta's Inappropriate Interactions with Children classifier only runs on material reported by users).

Microsoft	Xbox	By default	When an account was flagged, it was sent to the investigation team for deeper review. A full review of the account and its history was made and if the account was found to be engaged in grooming, the account and devices were permanently suspended and reported to NCMEC. The user can no longer log into the Xbox application or console. For personal computers and phones, Xbox also removes the ability of the user to access the Xbox application on their phone and personal computer.
Services that had a human moderator review only a proportion or none of the reports flagged by the language analysis technology			
Discord	Discord	Safety Alerts in Chats: All direct messages Sensitive Media Filter: In servers, direct messages and group direct messages	Notifications were sent to teenagers about potentially harmful messages, and they were offered safety information.
Google	YouTube	All material by default	The content was removed and action taken against the user in accordance with YouTube's policies, including potentially disabling a user's account.
Snap	Snapchat	Signals-based detection, Abusive language detection and Keyword-based Detection were used by default to prevent users from searching for terms, or creating usernames or display names, indicative of grooming. Signals-based detection and Keyword-based Detection were also used in chat, material and accounts that were reported to Snap by users or other individuals.	Snap disabled the account and took steps to prevent the account holder from creating a new one.

WhatsApp	WhatsApp Channels	By default	Internal proprietary tool A: If a suspected violation was detected, the relevant message, account, group or channel was sent for human review. Internal proprietary tool B: If a word or phrase associated with grooming was detected, it highlighted the word or phrase to facilitate human review.

Note 1: Apple's iCloud email, iMessage and FaceTime; Microsoft Teams; and Skype are not included in table 11 as they did not use language analysis technology to detect grooming.

Note 2: Discord stated that it reviewed a statistically significant sample of the communications identified by Safety Alerts in Chat feature to assess its effectiveness.

Note 3: Google stated that human moderators reviewed a subset of reports flagged by its language analysis technology to detect grooming on YouTube, but that the proportion of reports reviewed by human moderators was not available because YouTube's flags for human moderator review were not specific to the indicators flagged by language analysis technology to detect grooming.

Note 4: Refer to page 127 for alternative information provided by Meta in relation to the proportion of reports generated by automated tools which were reviewed by human moderators.

Note 5: Snap stated that approximately 76.9% of reports flagged by language analysis technology were reviewed by a human moderator.

Note 6: WhatsApp stated that its language analysis tools did not generate a binary flag as to whether a report was or was not likely to involve inappropriate interactions with a child (or any other violation type). Instead, these tools generated a score which reflected the likelihood that a report may violate one or more of WhatsApp's policies, which was then combined with signals (including signals from language analysis tools as well as others) to create an overall prioritisation score. The overall prioritisation score enabled WhatsApp to prioritise reports relating to high severity violations. There was no threshold or cut off score that automatically resulted in human review. The system works in a dynamic manner with all reports being enqueued in order of priority, based on the overall prioritisation score, and the order adjusted as new reports enqueued. WhatsApp was unable to isolate reports where language analysis tools contributed to the overall prioritisation score.

Proactively detecting CSEA activity – sexual extortion

Sexual extortion is a form of blackmail where someone threatens to release an intimate image or video of a person unless they comply with certain demands, such as financial payments. Sexual extortion of someone under 18-years-old is a form of child sexual exploitation and abuse.

Research from a recent joint eSafety and Australian Institute of Criminology survey of 1,953 adolescents aged 16 to 18 years living in in Australia found that more than 1 in 10 adolescents had experienced sexual extortion.[62] Of those who had experienced sexual extortion:

- 1 in 3 (32.3%) had experienced more than one instance of sexual extortion, and more than half (57.7%) had experienced sexual extortion before the age of 16
- 2 in 5 (41.4%) were extorted using digitally manipulated material

[62] Australian Institute of Criminology, 2025, Sexual extortion of Australian adolescents: Results from a national survey, accessed 26 March 2025, URL: https://www.aic.gov.au/publications/tandi/tandi712

- 2 in 3 (64.6%) were extorted by someone they had met online and had never met in person.

The majority of these reports – almost 1,200 – were from young people between 18 and 24 years and 90 per cent of reports were from males.

According to Thorn's research a common tactic used by perpetrators of sexual extortion is moving children to 'platforms that are less likely to detect the event and/or where the child may be more likely to share content'. [63] Thorn found that '65% of children had experienced someone attempting to get them to 'move from a public chat into a private conversation on a different platform.'[64]

For these reasons, eSafety asked providers about the tools they used in the detection and prevention of sexual extortion for both children and adults.

Use of language analysis technology to detect sexual extortion

As with livestreaming and grooming, language analysis tools can assign a probability to written messages (such as texts) or spoken conversations (such as chats) where terms, abbreviations, codes and hashtags associated with sexual extortion are being used, allowing the service to intervene.

Table 12 provides details on the language analysis tools that the providers were using to detect sexual extortion of adults and children.

[63] Thorn, 2024, 'Trends in Financial Sextortion: An investigation of sextortion reports in NCMEC CyberTipline data', accessed 9 July 2025, URL: https://www.thorn.org/research/library/financial-sextortion/
[64] Thorn, 2024, 'Trends in Financial Sextortion: An investigation of sextortion reports in NCMEC CyberTipline data', accessed 9 July 2025, URL: https://www.thorn.org/research/library/financial-sextortion/

Key insights:

- Apple's iMessage, FaceTime, iCloud email did not use language analysis tools to detect sexual extortion of adults or children.
- Discord did not use language analysis tools to detect sexual extortion of adults on any part of its service. It used language analysis tools to detect sexual extortion of children in direct messages but not on any other part of the service, including discoverable community servers, community servers or friend servers.
- Despite having and using language analysis tools on YouTube, Google did not detect sexual extortion of adults or children on Google Meet, Google Chat or Google Messages.
- Microsoft Teams did not use language analysis technologies to detect sexual extortion of adults or children.
- Skype did not use language analysis technologies to detect sexual extortion of adults or children.
- Snap used language analysis tools to detect sexual extortion of adults and children but only on material and accounts reported to Snap.[65]

Compared with previous responses

eSafety previously reported on the lack of tools used for detecting sexual extortion on Discord and Google services in our 2023 transparency report on CSEA.

While gaps remain, eSafety observed positive changes on some services:

In 2023, Discord reported that it did not use any tools to detect sexual extortion. It is therefore a positive improvement that Discord now reported using tools to detect sexual extortion of children in direct messages. Discord could continue to improve its practice by using tools to detect sexual extortion of children on more parts of its service, and by expanding the use of its tools to detect sexual extortion of adults on Discord.

[65] Snap used proactive detection to detect and block attempts to create usernames and display names indicative of sexual extortion as well as attempts to search for such terms.

Table 12: Language analysis technology used to detect likely terms, abbreviations, codes, and hashtags indicating sexual extortion

Provider	Service	Name of all language analysis technologies used to detect sexual extortion of adults	Name of all language analysis technologies used to detect sexual extortion of children
Discord	Direct messages	N/A	Safety Alerts in Chat[66]
Google	YouTube	BERT (Bidirectional Encoder Representations from Transformer) model	BERT
Meta	Facebook Facebook Messenger Threads	Internal proprietary tools	Internal proprietary tools
	Instagram	Internal proprietary tools	Instagram: Internal proprietary tools Instagram Direct: Internal proprietary tools
Microsoft	Xbox	Community Sift Internal proprietary tool A	Community Sift Internal proprietary tool A Internal proprietary tool B
Snap	Spotlight Discover Stories Chat Snaps	Signals-based detection Abusive language detection Keyword-based detection	Signals-based detection Abusive language detection Keyword-based detection
WhatsApp	WhatsApp (other than Channels)	XLMR	Internal proprietary tool A
	WhatsApp Channels	Internal proprietary tool A Internal proprietary tool B	Internal proprietary tool A Internal proprietary tool B

Note: Apple's iMessage, FaceTime and iCloud email; Discord's Discoverable community servers, Community servers and Friend servers; Google's Meet, Chat and Messages; Microsoft Teams; and Skype did not use language analysis technologies to detect sexual extortion of adults or children. Discord's direct messages did not use language analysis technologies to detect sexual extortion of adults.

[66] Discord stated that Safety Alerts in Chat was part of a new safety initiative, Teen Safety Assist, and was an experimental feature developed in collaboration with Thorn and launched to limited users in 2024. Safety Alerts in Chat aimed to identify messages that may be indicative of grooming behaviours. Upon detection, it notified a teen user about potentially harmful messages and offered a suite of safety tools, including options to block and report the sender, along with providing advice on dealing with an unwanted situation and encouraging the teen user to take a break from the conversation.

Languages analysed

It's important that providers have language analysis tools that operate in the languages of the communities that use their services. In Australia, the most common languages spoken in Australian homes, other than English, are Mandarin, Arabic, Vietnamese, Cantonese and Punjabi.[67]

Figure 3 provides a summary of the languages the tools operated in. For a full list of all languages please see the individual provider responses at the end of the report.

Key insights:

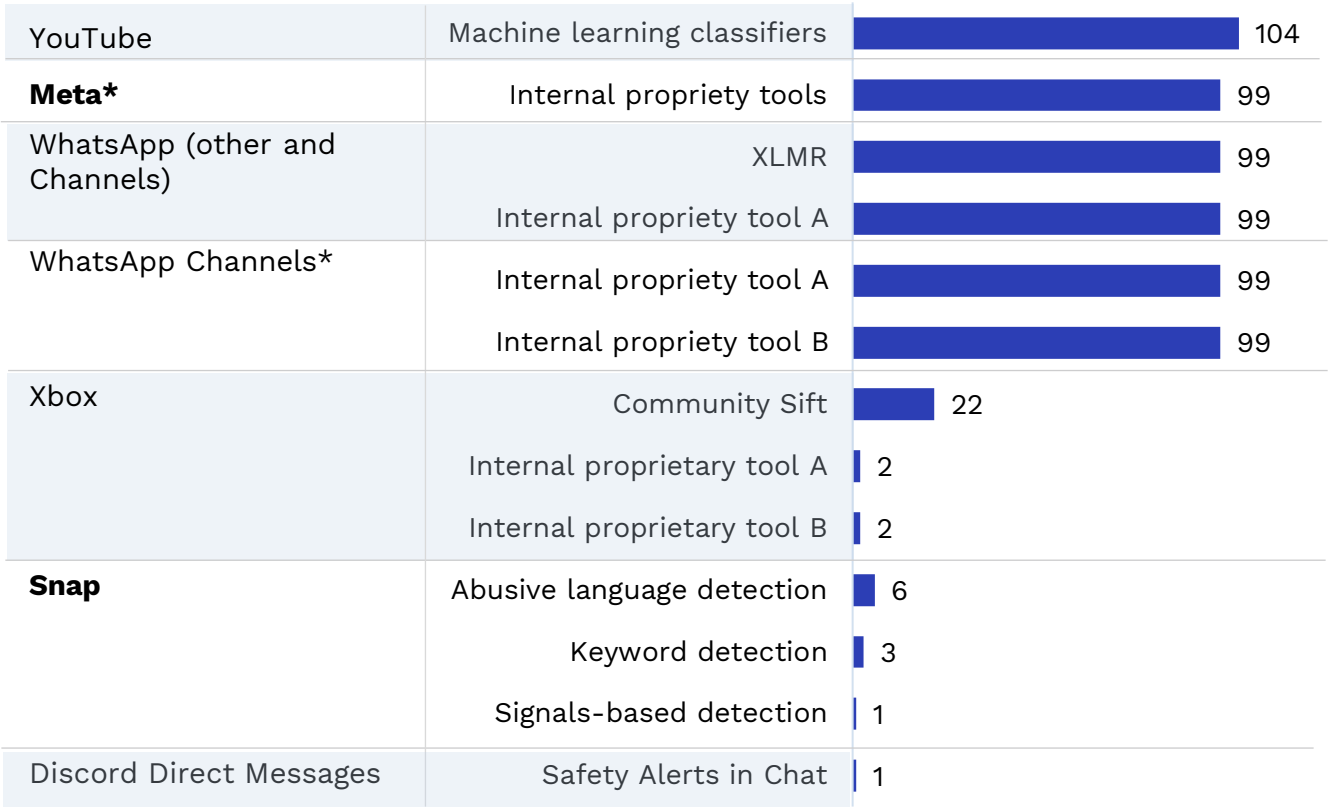
- Google's language analysis tools to detect sexual extortion reportedly operated in 104 languages, however the list of languages included the classical language Latin, alongside two invented languages (Volapük and Ido)[68] but did not include Cantonese[69] (see page 116 for the full list).
- Snap's language analysis tools to detect sexual extortion did not operate in any of the most common languages spoken in Australian homes other than English (Mandarin, Arabic, Vietnamese, Cantonese or Punjabi). These tools were only used on user reports, not proactively.
- Microsoft used multiple tools to detect sexual extortion of adults and children on Xbox, but two of these tools only covered two languages. The third tool covered 22 languages but did not include Punjabi or Cantonese (see page 135 for the full list).

[67] Australian Bureau of Statistics, 2022, 'Cultural diversity: Census', accessed 11 April 2025, URL: https://www.abs.gov.au/statistics/people/people-and-communities/cultural-diversity-census/latest-release#:~:text=Key%20statistics,-27.6%20per%20cent&text=61%2C860%20international%20visitors%20were%20in,Punjabi%20(0.9%20per%20cent).

[68] Invented (or constructed) languages are languages that are designed and created artificially, rather than evolving naturally through use. Invented languages are not typically spoken as native languages.

[69] Google's tools covered Chinese (Traditional) but not specifically Cantonese.

Figure 3: Number of languages used by analysis technologies to detect likely terms, abbreviations, codes and hashtags indicating sexual extortion



*Meta's Facebook, Facebook Messenger, Instagram and Threads, and WhatsApp Channels utilised an additional internal proprietary language agnostic tool. The languages covered by the tool depended on the languages of the words or phrases entered into the tool.

Note 1: Apple's iMessage, FaceTime and iCloud email; Discord's Discoverable Community Servers, Community Servers and Friend Servers; Google's Meet, Chat and Messages; Microsoft Teams; and Skype did not use language analysis technologies to detect sexual extortion of adults or children and are therefore not shown the in the table.

Note 2: Meta includes Facebook, Facebook Messenger, Instagram and Threads.

Note 3: Snap includes Stories, Spotlight, Discover, Chat and Snaps

As with grooming, there are a number of organisations that provide lists of terms, abbreviations, codes and hashtags that providers are able to utilise in detecting sexual extortion.

Table 13 provides details on the sources the providers were utilising.

Table 13: Organisations used to source terms, abbreviations, codes and hashtags indicating sexual extortion

Provider	All the organisations the service used to source terms, abbreviations, codes and hashtags
Discord	Thorn
Google	Google's own machine learning classifiers (Google's Content Safety API) [70]
Meta	Lantern, and Meta's own ongoing integrity work
Microsoft	Xbox: Lantern, NCMEC, Tech Coalition, Internet Watch Foundation and open-source intelligence.
Snap	Lantern
WhatsApp	None

Note: Apple, Microsoft Teams; and Skype did not utilise external organisations to source terms, abbreviations, codes and hashtags indicating sexual extortion and therefore not shown the in the table.

Steps services took if human moderators reviewed possible sexual extortion flagged by language analysis technologies

Providers were asked if human moderators reviewed all reports that were flagged by the language analysis technology. This is because harms such as sexual extortion are particularly context dependent and require understanding of language and culture. It is therefore particularly important that companies have human moderators operating in the languages of the communities they offer services to. Relying on language translation tools risks losing important context and cultural or linguistic nuance.

Table 14 shows the providers that did use human moderators for all or some of the sexual extortion reports that were flagged by language analysis technology, the language analysis technology that was used, whether the technology was used on all material by default and what steps were taken after sexual extortion was detected by the technology.

Consistent with other providers, Discord's notice required it to report the proportion of reports flagged by language analysis technology used to detect sexual extortion that were reviewed by human moderators. Discord stated that its human moderators reviewed a statistically significant sample of the communications identified by its safety measure, Safety Alerts in Chat, to assess its effectiveness but did not state the proportion of these reports flagged that were reviewed by human moderators.

[70] Google stated that YouTube used machine learning classifiers to help detect comments that may indicate child or adult sexual extortion and did not source specific terms or abbreviations from external organisations. Google stated that YouTube's machine learning models were trained to understand contextually what is being said, so as to more effectively detect child or adult sexual extortion activity.

Consistent with other providers, WhatsApp's notice required it to report the proportion of reports flagged by language analysis technology used to detect sexual extortion that were reviewed by human moderators. WhatsApp stated that its language analysis tools did not generate a binary flag as to whether a report was or was not likely to involve inappropriate interactions with a child (or any other violation type). Instead, these tools generated a score which reflected the likelihood that a report may violate one or more of WhatsApp's policies, which was then combined with signals (including signals from language analysis tools as well as others) to create an overall prioritisation score. The overall prioritisation score enabled WhatsApp to prioritise reports relating to high severity violations. There was no threshold or cut off score that automatically resulted in human review. The system works in a dynamic manner with all reports being enqueued in order of priority, based on the overall prioritisation score, and the order adjusted as new reports enqueued. WhatsApp was unable to isolate reports where language analysis tools contributed to the overall prioritisation score.

Table 14: Language analysis technology used and the steps taken when sexual extortion was detected and reviewed by a human moderator

Provider	Service	Name of the language analysis technology used and if used on all material by default	Steps taken when the language analysis technology detected sexual extortion on an account
Services that had a human moderator review all of the reports flagged by the language analysis technology			
Google	YouTube	Machine Learning classifiers (Google's Content Safety API) on all materials by default	The content was removed and action was taken against the user in accordance with YouTube's policies, including potentially disabling a user's account.
Meta[71]	Facebook Threads	Internal proprietary tool A: All material by default Internal proprietary tool B: Only used to help facilitate human review of user reports	For classifiers using internal proprietary tool A: Depending on signals and confidence of the classifier, content was either automatically deleted or

[71] Meta stated that, language analysis tools did not generate a binary flag as to whether a report was or was not likely to be sexual extortion or any other violation type, instead, these tools generated a score which reflected the likelihood that a report may violate one or more of Meta's policies, which was then combined with signals (including signals from language analysis tools as well as others) to create an overall prioritisation score. The overall prioritisation score enabled Meta to prioritise reports relating to high severity violations. There was no threshold or cut off score that automatically resulted in human review. The system worked in a dynamic manner with all reports being enqueued in order of priority, based on the overall prioritisation score, and the order adjusted as new reports enqueued. Meta was unable to isolate reports where language analysis tools have contributed to the overall prioritisation score.

	Facebook Messenger	Internal proprietary tool A: Only on Meta's public messaging products (Channels and Community Chats) and on material reported by users. Internal proprietary tool B: Only used to help facilitate human review of user reports	enqueued for human review. Meta may have also taken enforcement action at the account level. For classifiers using internal proprietary tool B: If the tool detected a word or phrase associated with sexual extortion, it highlighted the word or phrase to facilitate human review of the relevant material.
	Instagram	Instagram: Internal proprietary tool A: All material by default Internal proprietary tool B: Only used to help facilitate human review of user reports Instagram Direct: Internal proprietary tool A: Only on Meta's public messaging products (Channels and Community Chats) and on material reported by users. Internal proprietary tool B: Only used to help facilitate human review of user reports	
Microsoft	Xbox	Community Sift: All text areas by default Internal proprietary tool A: All text areas by default	When an account was flagged, it was reviewed by a human moderator to confirm that sexual extortion was occurring. If the human moderator confirmed this, the account and device of the user committing sexual extortion were terminated and removed from the service. Information about the user was submitted to NCMEC.
Services that had a human moderator review a proportion or none of the reports flagged by the language analysis technology			
Discord	Discord	Safety Alerts in Chat: All direct messages sent to teen users who were included in the trial during the reporting period	Notifications were sent to teenagers about potentially harmful messages and offer safety information
Snap[72]	Snapchat	Signals-based detection, Abusive language detection and Keyword-based Detection were used by default, to prevent users from searching for terms, or creating usernames or display names, indicative of sexual extortion.	Snap disabled the account and took steps to prevent the account holder from creating a new account.

[72] Approximately 66% of reports flagged by a language analysis technology were reviewed by a human moderator.

		Signals-based detection and Keyword-based Detection were also used in chats, material and accounts that were reported to Snap by users or other individuals.	
WhatsApp	WhatsApp (other than Channels)	XLMR: All reported material by default	XLMR: The relevant message, account, group, or channel was sent for human review. Internal proprietary tool B: If this tool detected a word or phrase associated with sexual extortion, it highlighted the word or phrase to facilitate human review.
	WhatsApp Channels	Internal proprietary tool B: All material by default	

Note 1: Apple's iCloud email, iMessage and FaceTime; Google's Meet, Chat and Messages; Microsoft Teams; and Skype are not included in table 14 as they did not use language analysis technology to detect sexual extortion of adults.

Note 2: Refer to page 127 for alternative information provided by Meta in relation to the proportion of reports generated by automated tools which were reviewed by human moderators.

Note 3: Discord stated that it reviewed a statistically significant sample of the communications identified by Safety Alerts to assess the effectiveness of the feature.

Note 4: WhatsApp stated that language analysis tools used by WhatsApp did not generate a binary flag as to whether a report was or was not likely to be sexual extortion (or any other violation type). Instead, these tools generated a score which reflected the likelihood that a report may violate one or more of WhatsApp's policies, which was then combined with signals from other tools to create an overall prioritisation score. The overall prioritisation score enabled WhatsApp to prioritise reports relating to high severity violations. There was no threshold or cut off score that automatically resulted in human review. Rather, the system operated in a dynamic manner with all reports being enqueued in order of priority, based on the overall prioritisation score, and the order adjusted as new reports were enqueued. WhatsApp was unable to isolate reports where language analysis tools contributed to the overall prioritisation score.

Proactively detecting CSEA in end-to-end encrypted services

End-to-end encryption (E2EE) can provide opportunities for various forms of covert online harms and distribution of illegal content, as many typical safety interventions can be harder to deploy. Of particular concern, it can allow offenders to have hidden contact with children, perpetrate exploitation and abuse of them, and store and share CSEA material. eSafety's position is that deployment of end-to-end encryption does not absolve services of responsibility for hosting or facilitating online abuse or the sharing of illegal content.

eSafety does not expect companies to design systemic vulnerabilities into services that use end-to-end encryption. Safety, privacy and security are not mutually exclusive, and each can be maintained through thoughtful and intentional design.

This approach is built into the Phase 1 Codes and Standards, with the Relevant Electronic Services and Designated Internet Services Standards requiring the detection of known CSEA material where technically feasible and reasonably proportionate. The standards also specifically provide that detection is not required where it would lead to systemic weaknesses

or vulnerabilities or render methods of encryption less effective. However, where these exemptions apply, a service must take appropriate alternative action.

Methods of proactively detecting known CSEA on the end-to-end encrypted services

Providers were asked if they undertook any work to investigate methods of detecting known CSEA on the end-to-end encrypted parts of their service(s).

Table 15 shows the providers who undertook such work and the details and aims of the development or research.

Table 15: Work that providers undertook to investigate methods of proactively detecting known CSEA on E2EE parts of their service

Provider	Service	Details	Aims of the research and/or development conducted
Apple	Apple services	Communication Safety in iMessage and FaceTime messages (end-to-end encrypted services) used an on-device machine-learning classifier to analyse photo and video attachments and determine whether a photo or video appeared to contain nudity.	Enable children to report nude photos and videos being sent to them in iMessage directly to Apple to inform its safety teams and products through Communication Safety. Create obstacles to storing or sharing **unlawful or harmful material** such as advising users through interstitials that the material might be unlawful or harmful or inappropriate, or referring to law enforcement. Create barriers to view or share sensitive content (such as interstitial warnings, blurring and blocking content, requiring a Screen Time passcode for the Family Sharing group to be entered on a child's device to view or send sensitive content through Communication Safety and Sensitive Content Warning). Limit a user's ability to share material with large numbers of people instantaneously.

Google	Google Messages	Google launched (to a closed group of beta users) Sensitive Content Warnings (SCW), an optional feature that was opt-out for users under 18 years of age. SCW blurred incoming images that may contain nudity before viewing. When an image that contains nudity was about to be shared or forwarded, it also reminded users of the risks of sending nude imagery.	Deliver meaningful on-device protection for all ages to prevent the sending and receiving of unwanted explicit nude imagery while using Messages.

Note: Meta's Facebook and Instagram; Snap's Snapchat; and WhatsApp did not undertake any work to investigate methods of detecting known CSEA on E2EE parts of their service and are, therefore, not included in table 15.

Time taken to respond to CSEA

Responding to user reports

The longer that CSEA is available on a service, the more likely it is to be accessed or seen by multiple users. This means the time taken to reach an outcome after CSEA is flagged is an important indicator of how effectively its distribution is disrupted, so immediate and ongoing harms to the victim-survivor are minimised.

The sooner online CSEA is reported to law enforcement, the sooner law enforcement officials can:

- start working to identify and safeguard the child, making sure they can access support services and other resources as appropriate[73]
- arrest the offender before they are able to abuse another child
- prevent further spread of CSEA by the offender and or any accomplices or people who have access to it.

[73] Know2protect, n.d., 'How2Report', accessed 11 July 2025, URL: https://www.dhs.gov/know2protect/how-to-report

Providers were asked about the median time it took to reach an outcome following potential CSEA material or activity being flagged by **automated tools** or reported by users, where the flags or reports were reviewed by human moderators. Human moderation is a key means of confirming that material or activity is actually CSEA-related.

Consistent with other providers, Apple's notice required it to report the median time taken for Apple Services to reach an outcome after receiving a user report relating to CSEA material, where those reports were reviewed by human moderators, as well as the absolute number of user reports. Apple stated that it was not able to provide this information as it was not tracking it during the relevant period. Apple stated that it was setting up a process to measure the resolution time for reports submitted to Apple.

Consistent with other providers, Google's notice required it to report the median time taken for YouTube to reach an outcome after receiving a user report relating to CSEA material, where reports were reviewed by human moderators. Google's notice also required it to report the median time taken for YouTube to reach an outcome after automated tools flagged potential CSEA, where those flags were reviewed by human moderators. Google stated that it was not able to provide this information as it did not have 'readily available data' to determine this metric. Google stated that 99% of all CSEA material on YouTube is proactively detected and most of this content was 'removed in minutes of being flagged' by YouTube's automated flagging systems.

Key insights:

- Services received a combined total of almost 11.5 million (11,458,969) user reports of CSEA globally and 374,261 reports from users in Australia.
- Compared to other providers, Meta services received the highest number of CSEA reports from users across Facebook, Instagram, Facebook Messenger and Threads, with almost 9 million reports (8,877,600) received globally during the report period and almost 40,000 (39,749) reports from Australian users. However, a high number of reports does not necessarily indicate that there is a higher volume of CSEA on these services in comparison to others. It could indicate that user reporting options were more easily accessible, or be influenced by the number of users on a services. In this case the reports received across the four Meta services have been aggregated which may also account for a larger number of reports than when compared to a provider of a single service

- Snap was the provider with the second highest number of user reports for CSEA, with 1,361,148 user reports globally and 27,796 reports from users in Australia from Snapchat.
- Meta's Threads responded the fastest to user reports of CSEA – the median time was 33 minutes, with a considerable number of reports (46,900).
- Compared to other services, Snap's human moderators had the fastest time to reach an outcome after automated tools flagged potential CSEA, with a median time of just 3 minutes.
- Apple's human moderators for iCloud email had the slowest median time to reach an outcome after automated tools flagged potential CSEA, taking a median time of 16 hours and 24 minutes.

Compared with previous responses[74]

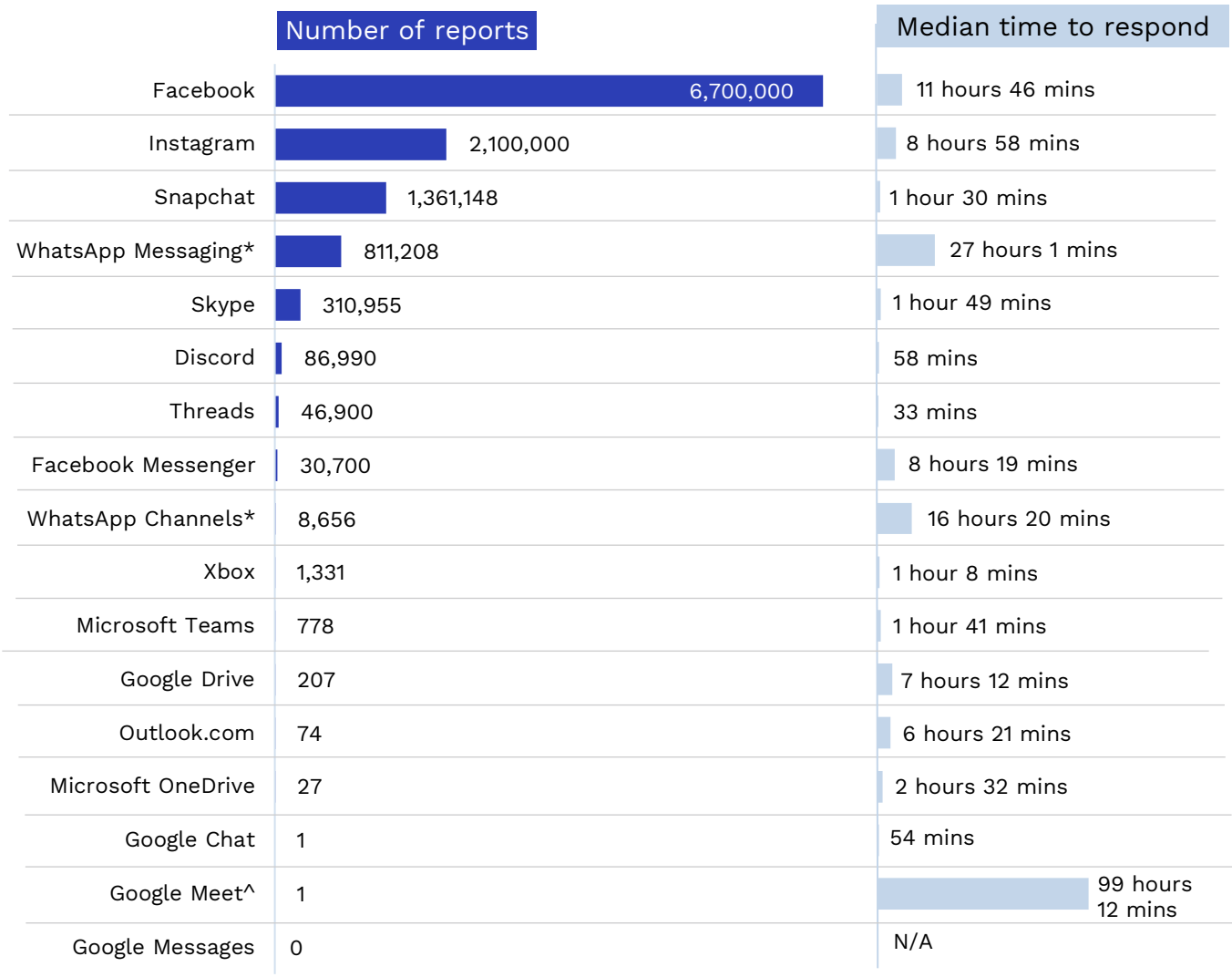
Some services improved the time taken to respond to reports of CSEA since they last received a notice from eSafety:

- Microsoft's median time to respond to user reports reduced significantly since 2022: from 1.5 days to 152.4 minutes for OneDrive, from 24 hours to 67.8 minutes for Xbox and from 3 days to 101.4 minutes for Teams.
- Skype's median time to respond to user reports reduced from 3 days to 109.2 minutes since 2022.

Figures 4 and 5 set out the number of reports received globally and in Australia in relation to each service, and the median time taken for the provider of the service to respond.

[74] Non-periodic notices were given in 2022 to Apple, Meta, Microsoft, Skype, Snap and WhatsApp, and in 2023 to Discord and Google. Note that the questions asked and the reporting timeframes differed between notice rounds and between providers. These comparisons serve as a guide to how service responses may have changed since the last notice they received. Future comparisons will be observed through the following three periodic notice responses from providers.

Figure 4: Number of user reports and median time taken to reach an outcome, where reports were reviewed by human moderators – global



*WhatsApp stated that as it did not categorise reports, this number was based on the number of accounts banned due to suspected CSEA activity, which had received a user report in the prior 30 days (to when they were banned).

^ Google stated that the global median response time for Meet reflected a single, specific, highly complex suspected grooming report and should not be taken as a typical turnaround time.

Note 1: Apple did not provide a response for iMessage, FaceTime, iCloud and iCloud email. Apple stated that it had been receiving user reports through a multi-channel reporting system, including in-service reporting tools, dedicated email aliases such as abuse@icloud.com and onlinesafetyau@apple.com, Apple Care, and Apple store retail associates. Apple stated that every suspected CSEA report, regardless of its source, received human review and is resolved as expeditiously as practicable. Apple stated that it was not tracking the information requested to answer this question using the specific metrics in the time period covered by this report, but that it is setting up a process to measure the resolution time for reports submitted to Apple.

Note 2: Discord stated that these metrics were calculated by evaluating received reports and corresponding outcomes for CSEA reports reviewed by human moderators for possible violations of Discord's policies. The time to report was defined by Discord as the time that passed between when a report was created to when the report received an action such as banning a response. Discord then calculated the median response time.

Note 3: Google stated that this metric was calculated from the time a user submitted a report to Google to when a content moderation decision was made by a human moderator within Google's review tool and an enforcement action was taken. Google stated that to the extent that users provide direct reports to Google through a non-child safety-specific queue, the median time it took Google's Child Safety team to action that content once it was flagged to the Child Safety team as suspected CSAM was between 30-60 minutes. **Google stated that it did not have 'readily available data' to determine the median time taken for YouTube to reach an outcome after receiving a user report relating to CSEA material, where reports were reviewed by human moderators. Google stated that 99% of all CSEA**

material on YouTube was proactively detected and most of this content was 'removed in minutes of being flagged' by YouTube's automated flagging systems.

Note 4: Meta calculated the the median time by identifying all user reports that were tagged as relating to potential violations of its child sexual exploitation, abuse and nudity policy (which may include violations relating to material that strictly falls outside the definition of CSEA as used in this report) and that were reviewed by human moderators and calculating the 50th percentile of the times taken from the submission of the user report to the time an outcome was reached by a human moderator. The data for Instagram also includes Instagram Direct.

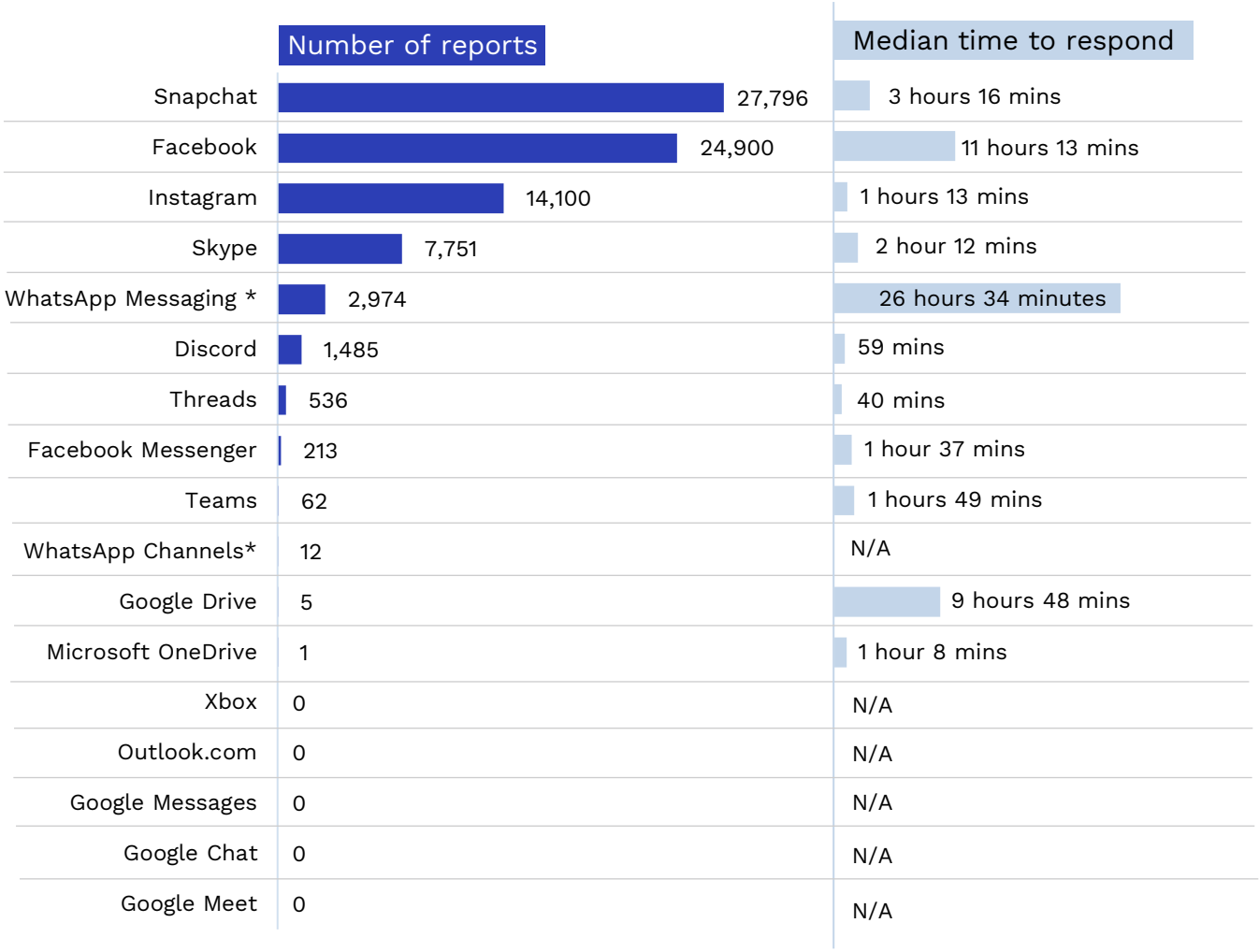
Note 5: Microsoft stated that the median time was calculated from the time a user reported a content item within in-application or centralised user reporting interfaces to when the human reviewer made a decision on the content item and the content item was actioned. Microsoft stated that most CSEA material was proactively detected through automated content detection (i.e., scanning), not through user reports. Microsoft's Digital Safety Content Report noted that, for the period January 2024 – June 2024, over 99% of all the content actioned for CSEA was detected through scanning.

Note 6: Skype stated that the median time was calculated from the time a user reported a content item in the user reporting interface and ending when the human reviewer made a decision on the content item and it was actioned. Skype stated that most CSEA material was proactively detected through automated content detection (i.e., scanning), not through user reports. Microsoft's Digital Safety Content Report stated that, for the period January 2024 – June 2024, over 99% of all the content actioned for CSEA was detected through scanning.

Note 7: Snap stated that that the number of reports was calculated using the unique number of user reports Snap received. The median response time was the 50th percentile of the total time between the time a report was created and the time a human moderator made a decision on the report. Global versus Australia data was determined using the reporting user's location.

Note 8: WhatsApp stated that it did not capture specific CSEA reporting categories. WhatsApp stated that Channels, the median time was calculated from the point when the material in that Channel is enqueued for human review to when an enforcement action is taken, where the enforcement action could be for a violation of any of WhatsApp's policies (not just those relating to CSEA). For accounts, the median time is from when a user report is enqueued for human review to when an enforcement action is taken, where the enforcement action is for a violation of WhatsApp's policies relating to CSEA specifically. WhatsApp stated that these accounts and channels had a user report in the previous 30 days (to when they were banned) and were banned after the first user report in the 30 day period. As the category of content reported for Channels is unknown, these metrics are only indicative of reported accounts for CSEA activity. WhatsApp stated that it did not log enforcement actions against specific user reports and so WhatsApp was not able to accurately calculate the median time taken to reach an outcome after receiving a user report. This metric was calculated based on the assumption that the maximum time for a report to be reviewed by WhatsApp's automated tools and enqueued for human review, was 24 hours. The 24 hours were then added to the median time taken once a report had been enqueued for review to when an enforcement action was taken.

Figure 5: Number of user reports and median time taken to reach an outcome, where reports were reviewed by human moderators - Australia



*As WhatsApp did not categorise reports, this number is based on the number of accounts banned due to suspected CSEA activity, which had received a user report in the prior 30 days (to when they were banned).

Note 1: Apple did not provide a response for iMessage, FaceTime, iCloud and iCloud email.

Note 2: See Note 2 on Figure 4 for Discord's response to how the figures were calculated.

Note 3: See Note 3 on Figure 4 for Google's response to how the figures were calculated.

Note 4: See Note 4 on Figure 4 for Meta's response to how the figures were calculated.

Note 5: See Note 5 on Figure 4 for Microsoft's response to how the figures were calculated.

Note 6: See Note 6 on Figure 4 for Skype's response to how the figures were calculated.

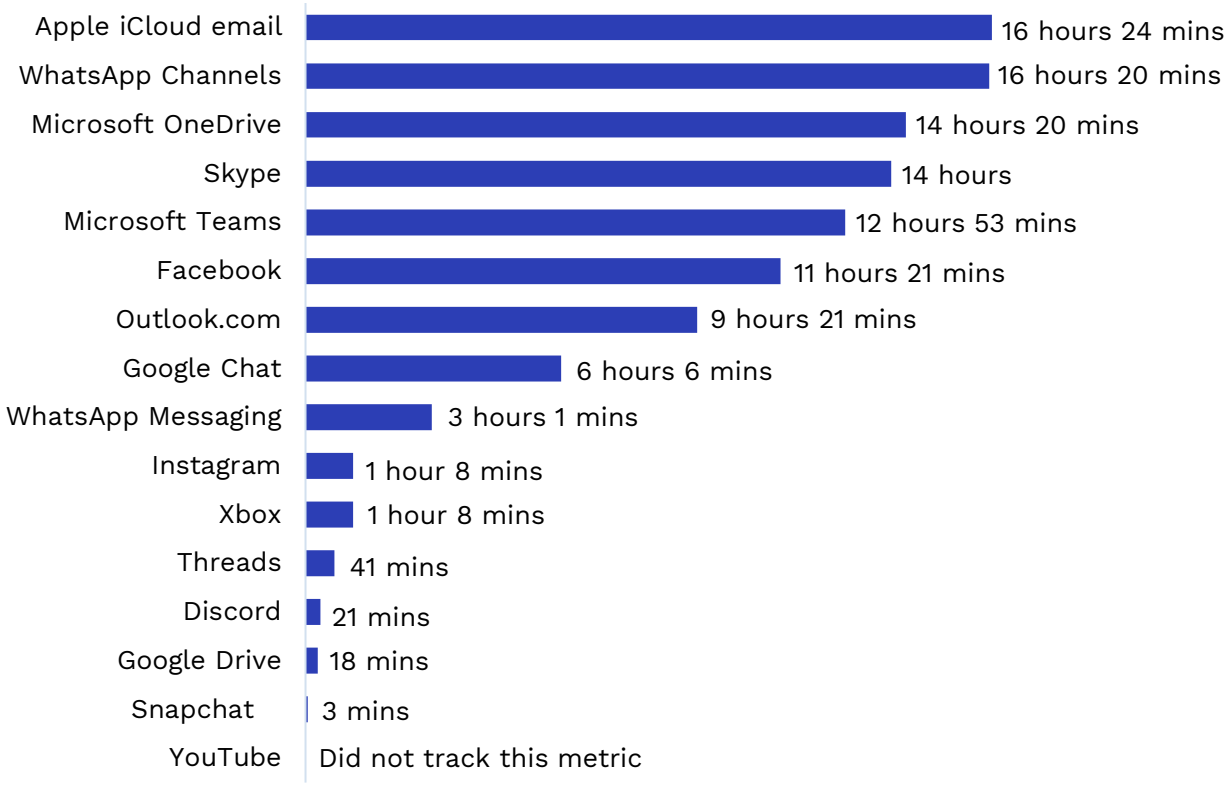
Note 7: See Note 7 on Figure 4 for Snap's response to how the figures were calculated.

Note 8: See Note 8 on Figure 4 for WhatsApp's response to how the figures were calculated.

Providers also used automated tools to flag potential CSEA, which were then reviewed by a human moderator to make a decision on whether the content was actually CSEA and if so, take an enforcement action.

Figure 6 shows the median time that was taken for providers to reach an outcome and action on material that was flagged by automated tools.

Figure 6: Median time taken to reach an outcome after automated tools flagged potential CSEA, where flags were reviewed by human moderators



Note 1: Apple stated that it did not scan for potential CSEA material on iMessage, Facetime and iCloud; Google stated that automated tools were not used on Meet and Messages; and Meta stated that automated tools were not used on Facebook Messenger. These services are, therefore, not included in Figure 6.

Note 2: Apple (iCloud email) stated that the median time was calculated by comparing (1) the timestamp of when an image matched against a known CSEA hash with (2) the corresponding case's 'last modified' status time stamp, as recorded in Apple's internal tool for managing CSEA-related incidents.

Note 3: Discord stated that it evaluated the reports created through hash-matching methods that were also reviewed by human moderators. The reported metric includes actions where Discord reviewers determine if the report did or did not violate Discord's policies.

Note 4: Google (YouTube, Drive, Meet, Chat, Google Messages and Gmail) stated that when apparent CSAM was identified using hash-matching technology or a classifier trained on previously identified CSAM, the potential CSAM content was enqueued for human review. This metric was calculated from the time of the initial identification of potential CSAM to when a content moderation decision was made by a human moderator within Google's review tool and an enforcement action was taken. Additionally, Google stated that to prevent repeated content exposure, Google automated the identification of those pieces of content for which Google had sufficient confidence based on past reviews. **Google stated it did not have 'readily available data' to determine the median time taken for YouTube to reach an outcome after automated tools flagged potential CSEA, where those flags were reviewed by human moderators. Google stated that 99% of all CSEA material on YouTube was proactively detected and most of this content was 'removed in minutes of being flagged' by YouTube's automated flagging systems.**

Note 5: Meta (Facebook, Facebook Messenger, Instagram and Threads) stated that the median time was calculated by material flagged by Meta's automated tools that was reviewed by human moderators and confirmed as violating its child sexual exploitation, abuse and nudity policy and calculating the 50th percentile of the times taken from the time at which the material was flagged to the time an outcome was reached by a human moderator.

Note 6: Microsoft (OneDrive, Outlook.com, Teams and Xbox) stated that the median time was calculated from the time that the referenced services flagged a submitted content item as potential CSEA imagery to the time that a human reviewer made a decision on the content item and the content item and/or account is actioned, which typically occurred within seconds following the human review decision.

Note 7: Skype stated that the median time was calculated from when Skype flagged a submitted content item as potential CSEA imagery to the time that a human reviewer made a decision on the content item and the content item was actioned, which Skype stated typically occurred within seconds following the human review decision.

Note 8: Snap stated that the median time was the 50th percentile of the total time between the time Snap's proactive models detected potential CSEA and the time a human moderator made a decision on the task created following proactive detection.

Note 9: WhatsApp stated that the median time was calculated by identifying all material flagged by automated tools as potentially violating its CSEA policies that was reviewed by human reviewers and calculating the median time from the time at which the material was enqueued for human review to the time an enforcement action was taken. It also noted that material flagged by the automated tools is enqueued for human review at, or very close to, the time of detection.

Availability of CSEA

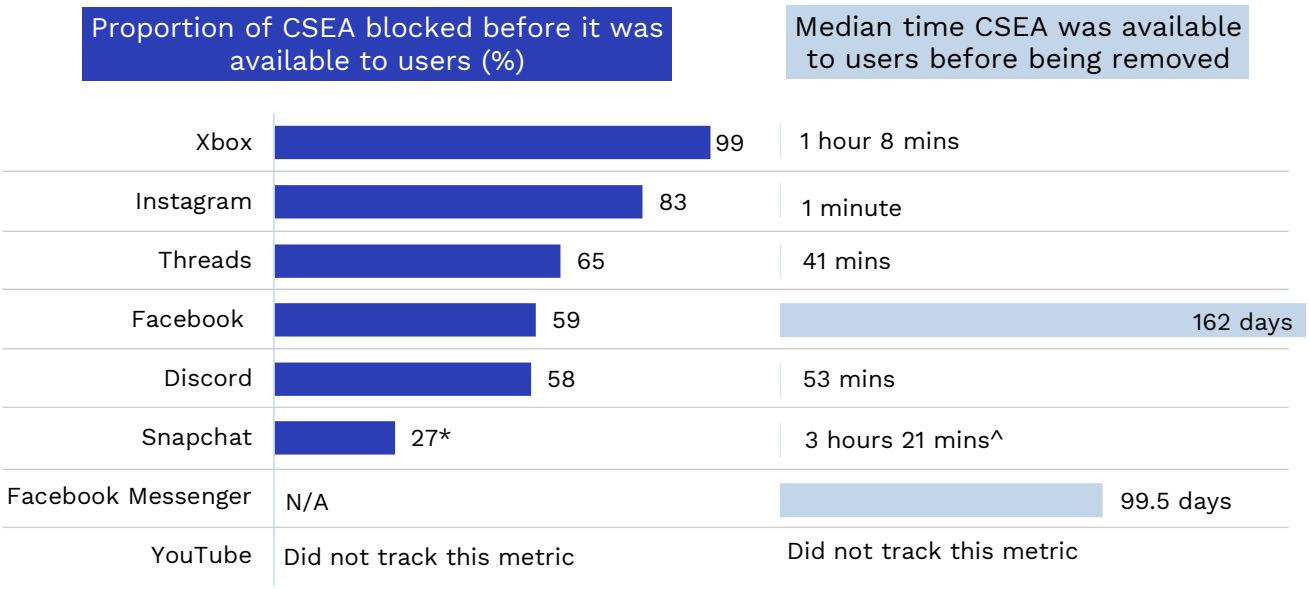
The longer that CSEA is available on a service, the further it can be distributed. Providers were asked questions about the median time that identified CSEA was available to users on the service and the proportion of CSEA material identified which was removed before it could be accessed by other users (see Figure 7). These are important indicators of the efficiency of providers' tools and the extent to which victim-survivors continued to be exploited through users viewing content depicting their abuse before it was detected and removed.

Google stated that it did not have 'readily available data' to determine the proportion of CSEA material identified and blocked/removed before the material was made available to other users on YouTube, nor the median time that CSEA material was available to users on YouTube to the time when it was removed. Google stated that 99% of all CSEA material on YouTube is proactively detected and most of this content was 'removed in minutes of being flagged' by YouTube's automated flagging systems.

Key insights:

- Instagram was quick to remove CSEA, reporting that material was available to users for 1 minute. Other services took up to an hour, or over an hour to remove CSEA, meaning it was available to other users for a substantial period of time.
- Compared to other services, Facebook and Facebook Messenger took longer to remove CSEA from their services. Meta reported that it took 233,600 minutes (162 days) to remove CSEA from Facebook and 143,300 minutes (99 days) to remove CSEA from Facebook Messenger. This is a stark contrast to the faster times displayed by another Meta service, Instagram.

Figure 7: Proportion of material identified and removed/blocked before available online (%) and the median time that CSEA was available to users on a service before it was removed.



*Snap stated that all content is available immediately once posted on Snapchat. Snap stated that there were cases in which Snap was able to take down content before it generated any views on Snapchat. Snap stated that it was only able to track these cases to the extent they related to content that was posted to Stories because it did not track the number of views for content other than Stories. To calculate the percentage of Stories containing CSEA that were blocked/removed before they were viewed by any other users on Snapchat relative to the total amount of CSEA material Snap identified, Snap calculated the total amount of CSEA Stories content that was taken down with 0 views and divided that by the total number of CSEA Stories content Snap enforced against in the reporting interval.

^Snap stated that this metric was only available for content that was posted to Stories on Snapchat. Snap stated that to calculate the median time, Snap took the 50th percentile of the total time it took between the time the Stories post was created and the time Snap took an enforcement action against that content for CSEA in the reporting interval.

Note 1: Discord calculated the median time that CSEA was available to users on the service to the time when it was removed by evaluating messages sent during the report period that were removed for violation of Discord's CSEA policies and were identified through automated tool and user reports. Discord stated that its hash tools processed applicable media before it was visible to other users. These tools accounted for over half of Discord's detected CSEA during the report period, which creates a median time availability of 0 minutes. Discord calculated the proportion of CSEA material identified and blocked/removed before the material was made available on the service by identifying messages sent during the report period that were deleted for CSEA at the time the user attempted to post the message. That number was then divided by the total number of CSEA messages removed from Discord during the reporting period.

Note 2: Meta (Facebook, Facebook Messenger, Instagram and Threads) stated that the median time was calculated by identifying all material that violated its child sexual exploitation, abuse and nudity policy and calculating the 50th percentile of the time taken from the creation of the material to the time at which it is removed. Meta stated that this includes data from proactive sweeps of historical content (7+ years) by its detection tools, which resulted in longer timeframes. The data for Instagram also included Instagram Direct (except to the extent that end-to-end encryption is enabled). The proportion of CSEA material identified and blocked/removed before material was made available to other users on the service was calculated by identifying all material that violated Meta's child sexual exploitation, abuse and nudity policy and calculating the percentage of such material that was removed before being viewed by another user. The data does not cover Facebook Messenger or Instagram Direct.

Note 3: Microsoft (OneDrive, Outlook.com, Teams and Xbox) reported that the median time was determined by comparing the upload time of all confirmed CSEA material to the time it was removed during the report period, then taking the median of those times. The proportion was calculated by taking the total number of confirmed CSEA material items blocked at the time of upload and dividing it by the total number of confirmed CSEA material items during the reporting period.

Note 4: Google stated that it did not have 'readily available data' to determine the proportion of CSEA material identified and blocked/removed before the material was made available to other users on YouTube, nor the median time that CSEA material was available to users on YouTube to the time when it was removed. Google stated that 99% of all CSEA material on YouTube was proactively detected and most of this content was 'removed in minutes of being flagged' by YouTube's automated flagging systems.

Preventing the creation and spread of CSEA

Limiting the availability of AI-generated CSEA

Generative AI can be misused to generate CSEA. This can include using generative AI to create deepfake sexually explicit images or videos, using text prompts or based on a regular image or video of a real child.[75]

In May 2025, NCMEC reported that its CyberTipline had seen 'a 1,325% increase in reports involving generative AI, going from 4,700 last year to 67,000 reports in 2024'.[76]

Even fake content can have a direct impact on the wellbeing and reputation of the victim. Threats to share it can also be used to blackmail them (a form of sexual extortion).

In addition, some generative AI tools can be used to impersonate human conversation convincingly and respond to humans in a highly personalised manner. This creates a risk that generative AI tools can be used to manipulate and abuse people, including children, for example through the creation of content to groom children.[77]

eSafety has also previously highlighted that harms from generative AI systems can occur unintentionally because of flaws in the data or models used to train the system, such as when abusive content is used for training.[78] To prevent the creation of CSEA, it is important that CSEA is excluded from training data-sets.

eSafety recognises 'Red-teaming' (testing to see if an AI system can be tricked into generating **harmful material**) is an important way of testing AI models to protect against harmful behaviour. Red-teaming may be conducted as part of developing a model, prior to launch or periodically after launch.

---

[75] National Centre for Missing & Exploited Children (NCMEC), 2024, 'CyberTipline 2023 Report', accessed 22 May 2025, URL: https://www.missingkids.org/cybertiplinedata, eSafety, 2023, 'Generative AI – position statement', accessed 22 May 2025, URL: https://www.esafety.gov.au/industry/tech-trends-and-challenges/generative-ai

[76] National Centre for Missing & Exploited Children (NCMEC), 2025, 'CyberTipline 2024 Report', accessed 26 May 2025, URL: https://www.missingkids.org/gethelpnow/cybertipline/cybertiplinedata

[77] National Center for Missing & Exploited Children, 2024, 'CyberTipline 2023 Report', accessed 22 May 2025, URL: https://www.missingkids.org/cybertiplinedata

[78] eSafety, 2023, 'Generative AI – position statement', accessed 22 May 2025, URL: https://www.esafety.gov.au/industry/tech-trends-and-challenges/generative-ai

Key insights:

Users of generative AI features and/or services on Discord, Gemini, Facebook, Facebook Messenger, Instagram and Snap could make in-service reports if their prompts generated CSEA material.

Among the providers that had integrated externally supplied third-party AI features on their services:

- Snap did not require assurances from its third-party providers that they undertook red-teaming on generative AI outputs specific to CSEA material.
- Discord and Snap did not conduct red-teaming of the generative AI feature outputs available via their services specific to CSEA material, citing legal barriers.
- Google and Meta, who both operated the generative AI models on which their generative AI services and/or features were based (that is, Gemini and Meta AI), used hash-matching tools (and, in the case of Google, machine learning classifiers) on all potential training data to detect and remove known and new CSEA images and videos before it was used as training data.

'In-service' reporting of generative AI

A select number of providers (Discord in respect of its third-party generative AI features; Google in respect of Gemini; Meta in respect of Meta AI on Facebook, Facebook Messenger and Instagram; and Snap in respect of Snapchat's AI Lenses, My AI, Dreams and AI Snap) were asked if users could make 'in-service' reports if their prompts using the services' generative AI features generated CSEA material (that is, without being required to locate a separate webform or email address on another website, or another part of the same site/app).

All of these providers provided 'in-service' reporting options if their prompts generated CSEA material.

Red-teaming on generative AI outputs

Providers (Apple in respect of its Apple Intelligence features; Discord in respect of its third-party generative AI features;, Google in respect of Gemini;, Meta in respect of Meta AIi on Facebook, Facebook Messenger and Instagram; and Snap in respect of Snapchat's AI Lenses, My AI, Dreams and AI Snap) were asked if they conducted their own 'in-house' red-teaming (simulation of misuse) of the generative AI model or feature outputs available via their services, specific to CSEA material. Providers were also asked if they conducted specific types of penetration testing as part of their red-teaming.

Table 16: Penetration testing features for providers who conducted red-teaming

Provider	Testing if the model(s) or feature(s) were capable of producing images/symbols associated with CSEA	Testing if the model(s) or feature(s) would refuse certain instructions, such as production of images/symbols associated with CSEA	Other testing
Apple	Yes	Yes	N/A
Google	Yes	Yes	Google conducted adversarial testing[79] of text responses to evaluate and address the model's ability to: advertise or solicit CSEA material generate graphic or sexualising stories or other text involving minors endanger minors such as supporting trafficking, enticement, or sexual extortion depict, encourage, or promote sexual attraction by adults towards children provide instructions to facilitate access of CSEA, abuse of minors, revictimisation of victim-survivors, evade detection of law enforcement.
Meta	Yes[80]	Yes[81]	N/A

Note: Discord and Snap did not conducting red-teaming of the generative AI model or feature outputs available via their services specific to CSEA material and therefore are not included in table 16.

[79] Google stated that as the term 'red teaming' does not have a uniform or consistent definition across industry and may encompass a wide array of activities, for the purposes of this response, Google referred to 'adversarial testing' rather than 'red-teaming'. Google stated that 'adversarial testing' means 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 within the limits of relevant laws (for example, without directly prompting a model to create CSAM). The process of adversarial testing undertaken for CSEA content is similar to that for other policy violating content. However, the adversarial testing for CSEA is subject to additional safeguards, and is only undertaken by specialist teams in a controlled and secure environment, to ensure it is compliant with applicable laws.

[80] Meta stated that existing US federal law made red-teaming exercises in this space more difficult, as it provided no immunity for such efforts. Meta stated that, nonetheless, Meta was working within the bounds of these laws to help ensure that its testing was as extensive as legally permissible.

[81] Meta stated that existing US federal law made red-teaming exercises in this space more difficult, as it provided no immunity for such efforts. Meta stated that, nonetheless, Meta was working within the bounds of these laws to help ensure that its testing was as extensive as legally permissible.

Providers who offered third-party generative AI features and/or models on their services were asked if they required assurance from third-party providers of generative AI features and/or models, that red-teaming had been undertaken on outputs specific to child sexual exploitation and abuse material.

Table 17: Assurances of red-teaming being undertaken on generative AI outputs from third-party providers

Provider	Service	Where were these assurances detailed to the provider (for example a contract, other agreement, public statements)? What did theses assurances consist cover?
Apple	Apple services	Apple contracted OpenAI to provide ChatGPT to Apple users. OpenAI has publicly disclosed its safety practices for its public AI models, which includes red-teaming[82]. Apple's contracts with third-parties typically included provisions requiring full compliance with applicable laws and regulations governing illegal and harmful content.
Discord	Discord	Discord's Developer Policy and Developer Terms of Service detailed requirements for developers that apps must not and are not used to violate Discord's Community Guidelines.[83]

Note: Snap's Snapchat -AI Lenses, My AI, Dreams and AI Snap did not have assurances from their third-party provider that red-teaming had been undertaken and therefore are not shown in table 17.

Providers were also asked about the frequency that they were performing red-teaming whether on proprietary generative AI or third-party generative AI features or models. Most providers (who were conducting red-teaming) were conducting testing prior to the launch of a new feature or model and some conducted red-teaming at varying intervals.

Table 18: Frequency at which red-teaming was performed

Providers	Frequency
Apple	Frequency was feature-dependent and based on significant model updates. For example, image generation evaluation/red-teaming was performed approximately twice a month.
Google	Google tested its generative AI products before launch and periodically after launch (for example when additional features or functionality was added to products or if the underlying model was retrained or updated). In addition to red-teaming, Google also deployed on-going monitoring tools to confirm that Gemini (after launch) was responding consistently with its policies.
Meta	Prior to the launch of a new model or changes to models (such as fine-tuning).

Note: Discord and Snap did not conduct red-teaming of the generative AI model or feature outputs available via their services specific to CSEA material and are therefore not included in table 18.

[82] OpenAI, 2025, 'Safety at every step', accessed 29 April 2025, URL: https://openai.com/safety/

[83] Discord stated that third-party apps were created and operated by developers and were made accessible to users on Discord using publicly available APIs. These third-party apps were not hosted on Discord, the developers were not Discord's service providers, and Discord did not have access to the generative AI models used by the developers. Users chose whether to add the apps to their servers / accounts and use them.

Providers were asked about the solutions they put in place to rectify any vulnerabilities identified by red-teaming.

Table 19: Solutions the provider put in place to rectify any vulnerabilities identified

Provider	Service	Solutions
Apple	Apple services	May have included safety models (checking input and/or output), keyword blocklists, and retraining models.
Google	Google services	Depending on the identified issues, Google may have made changes to address or correct the vulnerability identified before release. During the report period, Google implemented system instructions and supervised fine-tuning to mitigate a vulnerability identified during pre-launch adversarial testing.
Meta	Facebook – Meta AI Facebook Messenger - Meta AI Instagram – Meta AI	Filtering of training data to remove CSEA material from being used for training, fine-tuning of models to steer them away from creating violating content, improvements to classifiers to prevent requests from generating violating content.

Note: Discord and Snap did not conduct red-teaming of the generative AI model or feature outputs available via their services specific to CSEA material and are therefore not included in table 19.

Training datasets for generative AI features

Providers who offered their own proprietary generative AI features and/or models on their services were asked about the steps they took to ensure that **known** CSEA was not included in training data sets used for generative AI features.

Table 20: Steps Providers took to ensure known CSEA was not included in the training datasets and the tools used on their proprietary generative AI features and/or models

Provider	Feature	Did the provider take steps to ensure that known CSEA was not included in datasets used for generative AI features?	Did the Provider use hash-matching tools on all potential training data to detect and remove known CSEA images and videos before it was used as training data?	Name of tools used
Google	Gemini	Yes	Yes	CSAI Match and internal proprietary tools
Meta	Meta AI	Yes	Yes	Internal proprietary tool VMD5 PhotoDNA SimSearchNet++ PDQ

Providers who offer their own proprietary generative AI features and/or models on their services were also asked about the steps they took to ensure that **new** CSEA was not included in training data sets used for generative AI features.

Table 21: Steps Providers took to ensure new CSEA was not included in training datasets and the tools used on their proprietary generative AI features and/or models

Provider	Feature	Did the provider take steps to ensure that new CSEA material was not included in datasets used for generative AI features?	Did the Provider use automated tools on all potential training data to detect and remove new CSEA material before it was used as training data?	Name of tools used
Google	Gemini	Yes	Yes	CSAI Match and internal proprietary tools
Meta	Meta AI	Yes	Yes	CSAM content classifier CSAM solicitation content classifier Child sexualisation content classifier Google's Content Safety API

Providers which offered third-party generative AI features and/or models on their services were asked if they required assurances from providers of generative AI models for generative AI features available on their services to ensure that **known and new** CSEA material was excluded from training datasets.

Table 22: Assurances from third-party providers that generative AI models training datasets did not contain known or new CSEA material

Provider	Service	Known CSEA material		New CSEA material	
		Was assurance required?	If yes, did this process include requiring assurances that hash-matching tools were used on all potential training data to detect and remove known CSEA images and videos before they were used as training data?	Was assurance required?	If yes, did this process involve requiring assurances that automated tools were used on all potential training data to detect and remove new CSEA before it was used as training data?
Apple	Apple services	Yes	No	Yes	No
Discord	Discord	Yes	No	No	N/A
Snap	Snapchat - AI Lenses, My AI, Dreams and AI Snap	No	N/A	No	N/A

Preventing recidivism

'Recidivism' refers to banned or suspended users re-registering with a service using new details. eSafety investigators regularly see the same offenders create multiple new accounts, even after they have been banned by a platform. Preventing recidivism is an important safety measure to protect other users on a service from known offenders.

Recidivism can be detected and prevented by using signals or indicators of a banned user's account or online identity, to proactively prevent the same offender from re-registering.

Key insights:

- All services used indicators to prevent recidivism of users banned for CSEA violations.

Preventing recidivism relating to CSEA

All providers were using one or more indicators to prevent recidivism. eSafety has previously engaged with law enforcement agencies and child safety experts to seek views on what kind of information would not be in the public interest to publish and has determined it is not in the public interest to publish specific indicators or signals providers used.

Table 23 identifies where providers are using indicators for accounts banned or suspended for CSEA and the actions taken based on the indicators.

Table 23: Use of indicators and action taken on accounts banned or suspended to prevent recidivism relating to CSEA

Provider	Service	Were indicators used by default in all cases?	Specific action taken based on indicators
Apple	Apple services	All cases by default	Apple systems automatically blocked the creation of new accounts using indicators associated with previously banned accounts.
Discord	Discord	All cases by default	Where there was a high level of confidence, bans could be automatic. In other cases, signals were used as part of investigations into suspected users that might have previously been banned. If this was determined, the alternative account was also banned. For users suspected of attempting to circumvent a ban, Discord could force challenge via a CAPTCHA or phone text message verification. For high harm users, Discord could make this verification effectively permanent, requiring individuals to verify repeatedly without gaining access to the service.
Google	Google services	All cases by default	The content was removed and the account may have been disabled. When an account was disabled after finding CSEA, Google notified the user. Its message clearly stated that there was a violation of Google's policies and may be illegal. The message also linked to Google's disabled account help centre page, which described in detail why accounts are disabled for CSEA, informing users that Google reports child exploitation to NCMEC.
Meta[84]	Facebook Instagram	No[85]	For high-confidence indicators detected at account registration, the account was removed or prevented from being created.

[84] Meta stated that the action taken differed depending on the specific indicators and where in the life of any account the indicators were located.

[85] Meta stated that the set of indicators can vary depending on the account in question and the method of prevention. Meta did not have fixed criteria or thresholds for use of each indicator. Meta stated that given the adversarial nature of the issue, the set of indicators used changes over time, and Meta regularly reviews the data it uses to improve the performance of its solutions.

Microsoft	Microsoft services	All cases by default where an account had been banned or suspended for CSEA, or any other violation of Microsoft's terms of service	New account registrations were compared against accounts previously closed for CSEA using the indicators, to identify similarities with the CSEA-associated accounts and limit re-registration by the same account holders.
Skype	Skype	All cases by default where an account had been banned or suspended for CSEA, or any other violation of Microsoft's terms of service	New Microsoft account registrations were compared against accounts previously closed for Microsoft policy violations using indicators to limit re-registration by the same account holders.
Snap	Snapchat	By default, in all cases where an account was banned for CSEA	Based on indicators, attempts to re-register meant that new account registration would automatically fail.
WhatsApp	WhatsApp	All cases by default	Users were blocked from re-registering an account for a period of time.

Cross service information sharing to prevent recidivism relating to CSEA

Providers which have multiple services (Apple, Google, Meta and Microsoft) were asked about their processes to share information between services on indicators relating to accounts that had been banned or suspended for CSEA.

Table 24 identifies the providers that are sharing indicators between services and if they were used automatically to ban identified accounts.

Table 24: Cross service information sharing to prevent recidivism relating to CSEA

Provider	Services	Were indicators shared with all other provider services when an account on one of the provider's services was banned or suspended for CSEA?	Were these indicators used to automatically ban identified related accounts?
Apple	Apple services	Yes	Yes
Google	Google services	Yes	Yes
Meta[86]	Facebook Instagram	Yes Yes	Yes[87]
Microsoft	Microsoft services	Yes[88]	N/A

[86] Meta stated that it did not have a fixed criteria or thresholds for us of each indicator. Meta stated that given the adversarial nature of the space, the set of indicators Meta uses changes over time, and Meta regularly reviews the data it uses to improve the performance of its solutions.

[87] Meta stated that if a user's account on one Meta Service was disabled for a severe violation of Meta's child sexual exploitation, abuse and nudity policy, and Meta had a reliable signal that the user had an account on another Meta service, Meta would also disable that user's other account.

[88] User account information was maintained in the Microsoft account, not at the service level. Any enforcement actions taken against the Microsoft account were reflected in all connected Microsoft services.

Preventing recidivism for sexual extortion of children

Providers were asked about the indicators they were using to precent recidivism where an account has been banned or suspended for sexual extortion of children. All providers were using one or more indicators to prevent recidivism of sexual extortion of children, and also applying these indicators to prevent sexual extortion of adults.

Table 25 identifies where providers were using indicators and what actions were taken based on the indicators used to prevent recidivism when accounts were banned or suspended for sexual extortion of children.

Table 25: Indicators used and action taken to prevent recidivism relating to sexual extortion of children

Provider	Service	Were indicators used by default in all cases where an account has been banned or suspended for sexual extortion of children?	Specific action taken based on indicators used to prevent recidivism when accounts were banned or suspended for sexual extortion of children
Apple	Apple services	All cases by default	Apple systems automatically blocked the creation of new accounts using indicators associated with previously banned accounts.
Discord	Discord	All circumstances	Where there was a high level of confidence, bans could be automatic. In other cases, signals were used as part of investigations into suspected users that might have previously been banned. If this was confirmed, the alternative account was also banned. For users suspected of attempting to circumvent a ban, Discord could force challenge via a CAPTCHA or phone text message verification. For high harm users, Discord could make this verification effectively permanent, requiring individuals to verify repeatedly without gaining access to the service.
Google	Google services	By default in all cases	The user was banned and the content was removed. The user would not be able to open a new account based on the indicators. When an account was disabled after finding CSEA, Google notified the user. Its message clearly stated that this was a violation of Google's policies and may be illegal. The message also linked to Google's disabled account help centre page, which described in detail why accounts are disabled for CSEA, informing users that Google reports child exploitation to NCMEC.

Meta	Facebook Instagram	No[89]	For high-confidence indicators detected at account registration, the account was removed or prevented from being created. In most severe cases, Meta blocked a user from accessing the services from any account.
Microsoft	Microsoft services	All cases by default where an account had been banned or suspended for sexual extortion of children, or any other violation of Microsoft's terms of service	New account registrations were compared against accounts previously closed for sexual extortion of children using the indicators, to identify similarities with the CSEA-associated accounts and limit re-registration by the same account holders.
Skype	Skype	All cases by default where an account had been banned or suspended for sexual extortion of children, or any other violation of Microsoft's terms of use	New Microsoft account registrations were compared against accounts previously closed for Microsoft policy violations using the indicators to limit re-registration by the same account holders.
Snap	Snapchat	Yes	Based on indicators, attempts to re-register meant that new account registration would automatically fail.
WhatsApp	WhatsApp	All cases	Users were blocked from re-registering an account for a period of time.

Cross service information sharing relating to preventing recidivism for sexual extortion of children

Providers which have multiple services (Apple, Google, Meta and Microsoft) were also asked about their processes to share information between those services on indicators relating to accounts that had been banned or suspended for sexual extortion of children.

Table 26 identifies the providers that are sharing indicators between services and if they were used automatically to ban identified accounts.

[89] Meta stated that it did not have a fixed criteria or thresholds for use of each indicator. Meta stated that given the adversarial nature of this space, the set of indicators it uses changes over time, and Meta regularly reviews the data it uses to improve the performance of its solutions.

Table 26: Cross service information sharing to preventing recidivism relating to sexual extortion of children

Provider	Services	Were indicators shared with all other provider services when an account on one of the provider's services was banned or suspended for sexual extortion of children?	Were these indicators used to automatically ban identified related accounts?
Apple	Apple services	Yes	Yes
Google	Google services	Yes	No
Meta	Meta services	Yes[90]	Yes[91]
Microsoft	Microsoft services	Yes[92]	N/A

Preventing recidivism relating to sexual extortion of adults

Providers were asked if they were also using indicators to prevent recidivism for accounts banned for sexual extortion of adults.

All providers were using one or more indicators to prevent recidivism in relation to sexual extortion.

Table 27 identifies where providers are using indicators and the actions taken based on the indicators.

Table 27: Indicators used and action taken to prevent recidivism for sexual extortion of adults

Provider	Service	Were indicators used to prevent recidivism when accounts were banned or suspended for sexual extortion?
Apple	Apple services	Yes
Discord	Discord	Yes
Google	Google services	Yes
Meta	Facebook Instagram	Yes

[90] Meta stated that the specific indicators shared, and the circumstances in which they were shared, with other Meta Services depended on a number of factors, including availability of the indicators, Meta's confidence in them and the nature of the violation.
[91] Meta stated that if a user's account on one Meta Service was disabled for a severe violation of Meta's policies around sexual extortion of children, and Meta had a reliable signal that the user had an account on another Meta servicer, Meta would also disable that account.
[92] User account information was maintained in the Microsoft account, not at the service level. Any enforcement actions taken against the Microsoft account were reflected in all connected Microsoft services.

Microsoft	Microsoft services	Yes
Skype	Skype	Yes
Snap	Snapchat	Yes
WhatsApp	WhatsApp	Yes

Providers where then asked if they were using the indicators to prevent recidivism in all cases of sexual extortion against adults, or only in certain cases.

Table 28 identifies where providers were using indicators and what actions were taken based on the indicators used to prevent recidivism when accounts were banned or suspended for sexual extortion of adults.

Table 28: Preventing recidivism relating to sexual extortion of adults

Provider	Service	Were indicators used by default in all cases or only in certain circumstances?	Action taken based on indicators used to prevent recidivism when accounts were banned or suspended for sexual extortion of adults
Apple	Apple services	All circumstances	Apple systems automatically blocked the creation of new accounts using indicators associated with previously banned accounts.
Discord	Discord	All circumstances	Where there was a high level of confidence, bans could have been automatic. In other cases, signals were used as part of investigations into suspected users that might have previously been banned. If it was determined that the user had previously been banned, the alternative account was also banned. For users suspected of attempting to circumvent a ban, Discord could force challenge via a CAPTCHA or phone text message verification. For high harm users, Discord could make this verification effectively permanent, requiring individuals to verify repeatedly without gaining access to the service.

Google	Google services	All cases by default	For Drive, Gmail, Meet, Chat and YouTube, Google prevented new accounts from being created based on indicators, unless and until the restriction had been successfully appealed by the user.
Meta	Facebook Instagram	No[93]	For high-confidence indicators detected at account registration, the account was removed or prevented from created.
Microsoft	Microsoft services	All cases by default where an account had been banned for sexual extortion of adults, or any other violation of Microsoft's terms of service.	New Microsoft account registrations were compared against accounts previously closed for Microsoft policy violations using indictors, to limit re-registration by the same account holders.
Skype	Skype	All cases by default where an account had been banned or suspended for sexual extortion of adults, or any other violation of Microsoft's terms of service	New Microsoft account registrations were compared against accounts previously closed for Microsoft's policy violations using indicators, to limit re-registration by the same account holders.
Snap	Snapchat	By default, in all cases where an account was banned for sexual extortion of adults	Based on indicators, attempts to re-register meant that new account registration would automatically fail.
WhatsApp	WhatsApp	All cases	Users were blocked from re-registering an account for a period of time.

Cross service information sharing to prevent recidivism relating to sexual extortion of adults

Providers which have multiple services (Apple, Google, Meta and Microsoft) were asked about their processes to share information between these services on indicators relating to accounts that had been banned or suspended for sexual extortion of adults.

Table 29 identifies the providers that are sharing indicators between services and if they were used automatically to ban identified accounts.

[93] Meta stated that did not have a fixed criteria or thresholds for each indicator. Meta stated that given the adversarial nature of this space, the set of indicators Meta uses changes over time, and Meta regularly reviews the data it uses to improve the performance of its solutions.

Table 29: Cross service information sharing to prevent recidivism relating to sexual extortion of adults

Provider	Services	Were indicators shared with all others of the provider's services when an account on one of the provider's services was banned or suspended for sexual extortion of adults?	Were these indicators used to automatically ban identified related accounts?
Apple	Apple services	Yes	Yes
Google	Google services	Yes	No
Meta	Meta services	Yes[94]	Yes[95]
Microsoft	Microsoft services	Yes[96]	N/A

Testing recommender systems

eSafety has previously highlighted the risks that children in particular may face from certain content being amplified by recommender systems.[97] Recommender systems also have the potential to expose them to heightened risks of online sexual exploitation and abuse through friend and follower suggestions that could lead to children to interact with potentially dangerous adults.

Key insights:

- All the services stated that they tested their recommender systems to ensure that unlawful and harmful material and activity were not amplified.

Providers that use recommender systems (Google, Meta and Snap) were asked if their systems were tested to ensure that unlawful or harmful material and activity were not amplified.

Table 30 shows that all services were testing their recommender systems.

[94] Meta stated that the specific indicators shared, and the circumstances in which they were shared, with other Meta Services depended on a number of factors, including availability of the indicators, Meta's confidence in them and the nature of the violation.
[95] Meta stated that if a user's account on one Meta Service was disabled for a severe violation of Meta's policies around sexual extortion, and Meta had a reliable signal that the user had an account on another Meta servicer, Meta would also disable that account.
[96] User account information is maintained in the Microsoft account, not at the service level. Any enforcement actions taken against the Microsoft account were reflected in all connected Microsoft services.
[97] eSafety, 2022, 'Position statement: Recommender systems and algorithms', accessed 16 May 2025, URL: https://www.esafety.gov.au/industry/tech-trends-and-challenges/recommender-systems-and-algorithms

Table 30: Recommender system testing

Provider	Service	Were recommender systems tested to ensure that unlawful and harmful material and activity were not amplified?
Google	YouTube	Yes
Meta	Facebook Feed Instagram Feed Instagram Reels Threads 'For You' Feed	Yes
Snap	Snapchat	Yes

Sharing information to reduce CSEA (and sexual extortion of adults)

Consultation and cooperation across industry helps minimise CSEA material and activity online.

One mechanism that promotes cross-platform signal sharing is the Lantern program, set up with the intent 'for companies to strengthen how they enforce their child safety policies'.[98] It works by facilitating the sharing of signals of the activity and accounts that violate providers policies against CSEA. Signals can be, for example, information like email addresses, usernames, CSAM hashes, or keywords used to groom as well as buy and sell CSAM. [99]

In its 2024 Transparency Report, Lantern stated that during the period 1 January 2024 to 1 December 2024, 296,336 new signals were uploaded into Lantern, totalling over a million signals to combat CSEA.[100] Lantern's 2024 Transparency Report also highlighted that although some providers enrolled in the program, Discord, Snap and Meta were the only providers in scope of the Notice that met the program's engagement criteria by making regular signal contributions in 2024[101].

There are also organisations that help prevent people from becoming victims of sexual extortion and other forms of image-based abuse. Take it Down[102] is a global hash-matching service operated by the U.S.-based NCMEC which helps remove and prevent distribution of online nude, partially nude, or sexually explicit photos and videos of children under 18 (or that were taken before a person turned 18). The user can access a free tool to hash specific intimate images or videos of themselves, so attempts to upload them are blocked on a wide range of services.

---

[98] Tech Coalition, 2023, 'Announcing Lantern: The first child safety cross-platform signal sharing program', accessed 12 March 2025. URL: https://www.technologycoalition.org/newsroom/announcing-lantern.
[99] Tech Coalition, 2023, 'Announcing Lantern: The first child safety cross-platform signal sharing program', accessed 12 March 2025. URL: https://www.technologycoalition.org/newsroom/announcing-lantern.
[100] Tech Coalition, 2025, 'Lantern: Advancing child safety through signal sharing. Transparency Report 2024', accessed 17 July 2025, URL: https://www.technologycoalition.org/knowledge-hub/lantern-2024-transparency-report.
[101] Tech Coalition, 2025, 'Lantern: Advancing child safety through signal sharing. Transparency Report 2024', accessed 17 July 2025, URL: https://www.technologycoalition.org/knowledge-hub/lantern-2024-transparency-report.
[102] Take it down. 2025, 'Take It Down.', accessed 12 March 2025, URL https://takeitdown.ncmec.org.

StopNCII.org is a similar global service operated by UK-based Revenge Porn Helpline for adults (aged 18+) who are experiencing intimate image abuse.[103] Its technology is used by tech companies to help people from becoming victims by preventing the sharing of specific intimate images. In this way it offers victims a preventative tool.

Providers were asked if they were members of some of these initiatives for minimising child sexual exploitation and abuse activity and material and sexual exploitation of adults.

Tables 31 and 32 show the providers who are part of these programs and how they are utilising the services.

Key insights:

- Despite being a member of Lantern, Google neither contributed signals to nor ingested signals from Lantern to address CSEA on Google Services.
- Skype was not a member of any cross-industry initiatives to reduce CSEA or the sexual extortion of adults.
- WhatsApp was not a member of any cross-industry initiatives to reduce CSEA or the sexual extortion of adults.

[103] StopNCII.org, 2025, 'About us', accessed 12 March 2025, URL: https://stopncii.org/about-us/..

Initiative membership

Table 31: Providers who were members of the Lantern program and how they were utilising the service

Provider	Steps taken as a result of ingesting signals taken from Lantern
Discord	Relied on signals to support the investigation and removal or user accounts. The signals may have also led to the removal of content that violates Discord's Community Guidelines.[104]
Google	None. Despite being a member of the initiative, Google stated that it did not share signals with Lantern or take signals from Lantern during the report period.
Meta	Automated tools checked signals in Lantern database against signals on Meta services. Automated tools then routed the information to the correct investigation team. An investigation was conducted on the service, and if harmful behaviour was confirmed, then Meta took action in line with its policies, which may have included, removal of content, application of a strike to the account or restricting or disabling the account.
Microsoft[105]	Xbox is the only service that was ingesting signals from Lantern. When a signal matched an account on Xbox's service, the account was further investigated. If evidence of CSEA was found, the account was actioned and any CSEA content was reported to NCMEC. If there was no evidence of CSEA, the flagged account continued to be monitored for CSEA content.
Snap	Snap may have investigated the account and, if the user was determined to have engaged in violations of any of Snap's Terms of Service (including Snap's Community Guidelines), Snap took appropriate enforcement action in accordance with policies. Snap used URLs ingested from Lantern to prevent users from posting content containing the URLs. Snap used keywords shared via Lantern to: (1) supplement its understanding of how violating behaviour may manifest on the platform and (2) prevent users from searching for the terms or creating display names and / or usernames containing them by ingesting these keywords into its Abusive Language Detection system.

Note: Apple, Skype and WhatsApp were not members of Lantern.

[104] Discord was a founding member of Lantern and continues to be an active participant in working groups within the Lantern initiative.
[105] Xbox was the only Microsoft service that is a member of the Lantern program.

Table 32 shows other initiatives that providers are utilising such as Take it Down or StopNCII.

Table 32: Other programs or relevant initiatives that providers were associated with

Provider	Other relevant initiatives
Apple	Apple received hashes from Cybetip.ca, IWF and NCMEC. Apple sent validated reports of CSAM images and videos to NCMEC.
Discord	IWF SWGfl (StopNCII) Revenge porn helpline Tech Coalition INHOPE EU Internet Forum Safety Reporting Network[106]
Google	Safety By Design Generative AI principles (developed by Thorn and All Tech is Human) U.S. Department of Homeland Security's Know2Protect campaign NCMEC's No Escape Room initiative
Meta	Take It Down StopNCII Tech Coalition WePROTECT
Microsoft	Tech Coalition NCMEC StopNCII.org Gaining intelligence, as well as industry and civil society perspectives that inform Microsoft policies and approach.[107]
Snap	Take It Down StopNCII Report Remove

Note 1: Skype and WhatsApp were not members of any other initiative and therefore are not shown in table 32.

Note 2: Skype stated that it benefits from Microsoft's safety partnerships, such as the Tech Coalition and National Center for Missing and Exploited Children, where it receives valuable intelligence, as well as industry and civil society perspectives that inform Microsoft policies and approach.

[106] Discord stated that it also operated its voluntary Safety Reporting Network, allowing it to collaborate with various organisations around the world to identify and report violations of its Community Guidelines. Members of Discord's Safety Reporting Network have access to a prioritised reporting channel. Discord partners with many organizations through the Network, including Australian partners such as Kids Helpline and the eSafety Commissioner.

[107] Microsoft stated that all Microsoft services benefit from Microsoft's safety partnerships as it received valuable intelligence, as well as industry and civil society perspectives that informed Microsoft policies and approach.

Resourcing to ensure safety

The number of Trust and Safety staff, along with the language skill set of moderators, can impact the ability of providers to address CSEA material and activity.

According to the 5Rights Foundation (an UK-based NGO):

> '[Trust and Safety staff] plays a crucial role in defining and implementing safety standards. Their responsibilities include writing platform policies (both front-facing policies such as Community Guidelines and the more detailed policies that content moderators apply in their day-to-day work); adjudicating 'edge case' content and issues (escalations) that require specialist review by senior experts; developing and deploying the machines that proactively detect potentially violative content, contact, and conduct; responding to requests from and/or reporting incidents to national and international law enforcement and security services; resourcing and managing the work of content moderation teams; collecting and analysing data relevant to safety (metrics); developing safety tools and resources (for example parental controls or safety hubs); engaging with external experts (for example academics and NGOs); advising on the levels of safety of new product and service proposals; supporting compliance; and transparency reporting.[108]
>
> Platforms with larger, more mature [Trust and Safety] functions may also include engineering teams that build and maintain moderation tools, training and knowledge management for moderators, researchers and data analysts who provide the evidence base that informs all aspects of [Trust and Safety] work including the level and prevalence of harm, and intelligence and discovery teams to detect emerging threats. The law enforcement response and compliance team handle legal requests and ensure regulatory compliance (often in collaboration with Legal).'[109]

[108] Shulru, T., 2024, 'Trust and Safety work: internal governance of technology risks and harms.', Journal of Integrated Global STEM, Vol. 1 (Issue 2), pp. 95-105., accessed 17 July 2025, URL: https://doi.org/10.1515/jigs-2024-0003. See also Lehane, C., 2022), 'A Career in Trust & Safety: You know more than you know', accessed 17 July 2025, URL: https://medium.com/@christinemlehane/a-career-in-trust-safety-you-know-more-than-you-know-3a02f63059a4.

[109] 5Rights Foundation, 'Advancing Trust & Safety: systems and standards for online safety professionals', accessed 14 March 2025, URL: https://5rightsfoundation.com/resource/advancing-trust-safety-systems-and-standards-for-online-safety-professionals/

Number of staff

Providers were asked about the number of employees they had in these roles that were vital in keeping services safe.

Table 33 sets out the number of staff employed or contracted as Trust and Safety staff and Table 34 sets out the number of staff employed or contracted as content moderators.

Consistent with other providers, Apple's notice required it to report a categorised breakdown of its Trust and Safety staff. Apple did not provide all required information.

Consistent with other providers, Google's notice required it to report a categorised breakdown of its Trust and Safety staff. Google did not provide this information.

Table 33: Number of staff employed or contracted by providers as Trust and Safety staff including Engineers (as at 15 December 2024)

Provider	Engineers employed by provider focussed on Trust and Safety	Trust and safety staff employed by provider (other than engineers and content moderators)
Apple[110]	Did not provide a response	Did not provide a response
Discord	78	83
Google[111]	Did not provide a response	Did not provide a response
Meta[112]	2110	3915[113]
Microsoft	5 (working across all Microsoft services except Xbox) 8 additional for OneDrive 38 additional for Xbox	30 across all Microsoft services, 3 additional for all Microsoft services (except Xbox) 1 additional Product manager for OneDrive 3 additional Trust and Safety Product Managers for Teams 165 staff for Xbox in varying roles

[110] Apple stated that it did not have individual category data to provide. Apple stated that it embeds responsibility for user safety (including child safety) across product teams throughout the company.
[111] Google stated that it did not have individual category data to provide.
[112] Meta provided data as at 30 September 2024 due to its quarterly record keeping. Meta will provide responses for all subsequent reports as at the end of the report period. Meta provided approximations for all metrics. Meta's data does not include staff focused on WhatsApp. Meta stated that the data does not represent the full spectrum of people working on Trust and Safety at Meta. Rather, the data represents teams focused on core trust and safety work for Meta. Meta stated that it also had people working on trust and safety who were distributed globally and within multiple different teams that work on some aspects of trust and safety to maintain the safety and security of Meta's services.
[113] Meta stated that these refer to Trust and Safety staff that were not engineers. This group includes employees working in global operations, other non-engineering tech functions (for example product managers, researchers, designers etc.), legal and policy. Meta stated that it also had thousands of contractors (other than content moderators) who support its Trust and Safety work.

Skype	5 engineers employed by Microsoft (who work across several Microsoft services)	3 staff dedicated to Skype Additional 30 trust and safety employed by Microsoft (working across several Microsoft services)
Snap	98	97[114]
WhatsApp[115]	157	235[116]

Table 34: Number of staff employed or contracted by providers as Content Moderators (as at 15 December 2024)

Provider	Content moderators employed by provider	Content moderators contracted by provider	Any other information
Apple[117]	Over 500	Did not provide a response	N/A
Discord	35	246	N/A
Google[118]	Did not provide a response	Did not provide a response	N/A
Meta[119]	0[120]	31,443	N/A
Microsoft	29 across all Microsoft services	41 (working across all Microsoft services except Xbox) 15 content moderators and 30 investigators for Xbox	In addition, all services except Xbox leveraged support from a group of 68 vendors contracted by Microsoft on a centralised team that provided other trust and safety-related support such as account appeals, government requests, and internal escalations.

[114] Snap stated that the data it provided for the number of Trust and safety staff employed (other than engineers and content moderators) reflected staff employed by Snap on the following teams: Law Enforcement Operations, Identity Operations, Community Support, Platform Policy, Platform Safety, Safety Legal, Safety Product and Safety Strategy & Programs teams.

[115] WhatsApp provided data as at 30 September 2024 due to its quarterly record keeping. WhatsApp will provide responses for all subsequent reports as at the end of the report period. WhatsApp stated that the figures provided represent teams who were focused on core trust and safety work for WhatsApp. WhatsApp had additional Trust and Safety staff who work on some aspect of trust and safety to maintain the safety and security of the service.

[116] This group included employees working in global operations and other non-engineering tech functions (for example, product managers, researchers, designers, etc.)

[117] Apple stated that it did not have individual category data to provide. Apple stated that it embeds responsibility for user safety (including child safety) across product teams throughout the company.

[118] Google stated that it did not have individual category data to provide.

[119] Meta provided data as at 30 September 2024 due to its quarterly record keeping. Meta will provide responses for all subsequent reports as at the end of the report period. Meta provided approximations for all metrics. Meta's data does not include staff focused on WhatsApp. Meta stated that the data does not represent the full spectrum of people working on Trust and Safety at Meta. Rather, the data represents teams focused on core trust and safety work for Meta. Meta stated that it also had people working on trust and safety who were distributed globally and within multiple different teams that work on some aspects of trust and safety to maintain the safety and security of Meta's services.

[120] Meta stated that content moderators were generally employed by Meta's vendors. However as at 30 September 2024, Meta had approximately 1935 employees in its global operations team, which focuses on work related to content moderation (for example quality reviews, content escalations, building protocols, managing contractors, etc.).

Skype	29 content moderators employed by Microsoft (who work across several Microsoft services)	41 vendors Contracted by Microsoft (who work across several Microsoft services)	68 vendors contracted by Microsoft that provided other trust and safety-related support such as account appeals, government requests and internal escalations.
Snap	89	1803[121]	
WhatsApp [122]	0[123]	1,407	

Number of languages covered

Review by human moderators may be required to verify CSEA flagged by tools or reported by users, particularly where the material has not been identified previously, or the activity is being livestreamed, or grooming or sexual extortion is being attempted. Effective moderation depends on understanding the language and culture for context, so it's important that providers have human moderators operating in the languages of the communities that use their services. Some services may use translation services when reviewing material or activity involving languages that their human moderators do not cover. However, eSafety does not consider automated translation services as a substitute for human moderation because relying on language translation tools risks losing important context and cultural or linguistic nuance. The most common languages spoken in Australian homes, other than English are Mandarin, Arabic, Vietnamese, Cantonese and Punjabi. Not all providers operate across these 5 languages, let alone other major languages spoken in Australia and by the providers' wider users.

Key insights:

- Apple's employed and contracted moderators only covered six languages. Apple did not have moderators who spoke Arabic, Punjabi, Cantonese or Vietnamese – four of the five most common languages spoken at home in Australia, apart from English.
- Discord did not have moderators operating in Punjabi or Cantonese.
- WhatsApp's moderators only covered 10 languages and did not include Punjabi, Cantonese or Vietnamese.

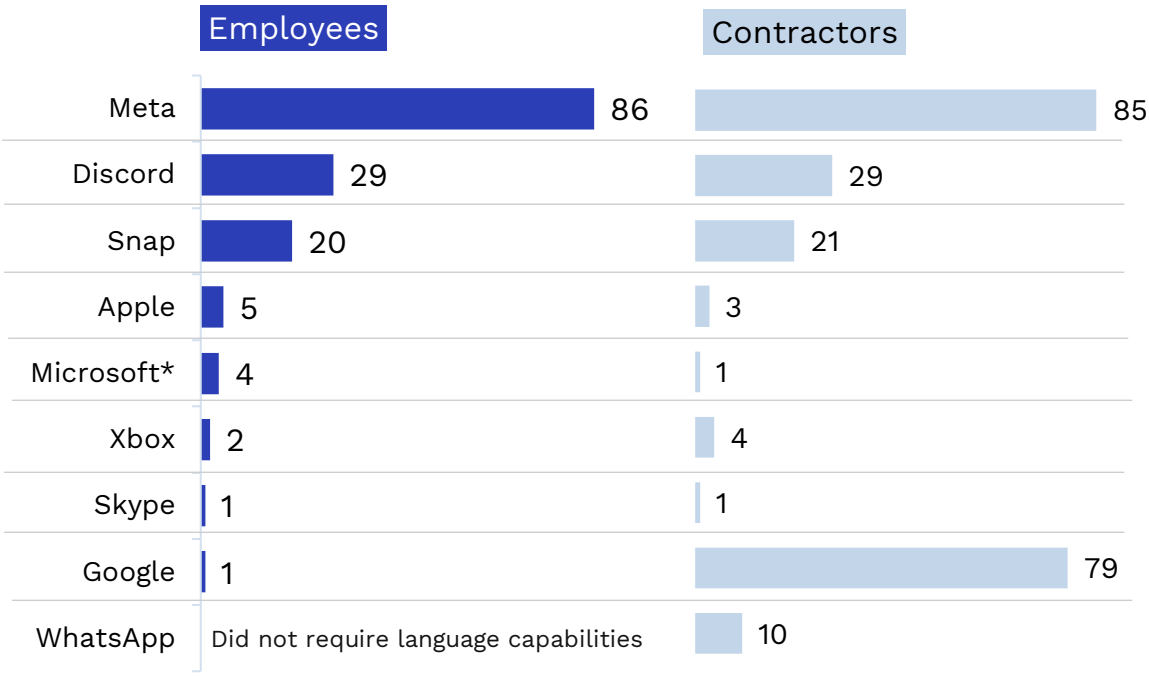
[121] Snap stated that it additionally contracts 92 law enforcement operations vendors.
[122] WhatsApp provided data as at 30 September 2024 due to its quarterly record keeping. WhatsApp will provide responses for all subsequent reports as at the end of the report period. WhatsApp stated that the figures provided represent teams who were focused on core trust and safety work for WhatsApp. WhatsApp had additional Trust and Safety staff who work on some aspect of trust and safety to maintain the safety and security of the service.
[123] WhatsApp stated that content moderators were generally employed by WhatsApp's vendors. WhatsApp had 203 employees in its global operations team focusing on human review processes (for example quality reviews, content escalations, building protocols, managing contractors, etc.).

Compared with previous responses[124]

Google reported that its content moderators covered more languages than in 2023 – increasing from 71 languages to 79.

Figure 8 sets out the number of languages spoken by providers' human moderators.

Figure 8: Number of languages that Provider's human moderators operated in



*Includes OneDrive, Outlook and Teams (excludes Xbox).

Note: WhatsApp did not track or require any specific language capabilities for trust and safety employees. WhatsApp stated that it relied on the language capabilities of its human review teams who were contractors.

[124] Non-periodic notices were given in 2022 to Apple, Meta, Microsoft, Skype, Snap and WhatsApp, and in 2023 to Discord and Google. Note that the questions asked and the reporting timeframes differed between notice rounds and between providers. These comparisons serve as a guide to how service responses may have changed since the last notice they received. Future comparisons will be observed through the following three periodic notice responses from providers.

Additional information from individual providers

Individual providers were also asked a series of questions specific to their services, or provided additional information to the tables and graphs in the section above.

Apple

Key insights:

- Apple did not have in-service reporting options on iCloud email, iCloud or FaceTime.
- Apple's iMessage and FaceTime are end-to-end encrypted. Apple relied on Communication Safety as a means of identifying CSEA on these services.

Compared with previous response[125]

In our 2022 transparency report, eSafety reported that Apple did not have **in-service user reporting** options within iMessage, FaceTime, iCloud or iCloud email. Since then, Apple has introduced in-service reporting in iMessage but only when Communication Safety is enabled and in-service reporting is still not available on its other services.

Proactive detection of CSEA

Was work undertaken to research, develop, source or implement tools to detect known CSEA images on iMessage, FaceTime and iCloud?

Apple responded 'yes' to this question.

Provide details and specify the aims of the research and/or development conducted to detect known CSEA images on iMessage, FaceTime and iCloud

Apple stated that it had implemented several child safety initiatives aimed at protecting its users and keeping its services safe including:

[125] A non-periodic notice was given to Apple in 2022. Note that the questions asked and the reporting timeframes differed between notice rounds. These comparisons serve as a guide to how service responses may have changed since the last notice they received. Future comparisons will be observed through the following three periodic notice responses.

- Released in September 2024 on iOS 18, iPadOS 18, and macOS Sequoia, Communication Safety required that the Screen Time passcode for the Family Sharing group be entered onto a child's device in order to view or send a sensitive image or video. The feature was developed with the aim of keeping children safe from inappropriate content and providing tools to parents to keep their children safe from harmful content including CSEA material.
- In December 2024, Apple also enhanced Communication Safety in iMessage by allowing users in Australia to report nudity flagged by Communication Safety directly to Apple. This was developed and designed to keep children safe, and all users safe from inappropriate content and provide users with tools to report CSEA.
- Apple used perceptual hashing technology in iCloud Mail to detect known CSEA material. Apple utilized this technology for outgoing emails for several years, and extended it in Australia to also detect known CSEA material in inbound emails beginning in December 2024.
- Since March 2004, Apple has provided research support for the Frontiers in Digital Child Safety initiative at the Technical University of Munich. Apple stated this is representative of its ongoing commitment to research and improving the safety of its services while preserving privacy. Since its launch, this interdisciplinary initiative has brought together experts from diverse fields such as social science, law, computer science, and engineering to collaborate on enhancing both existing and emerging approaches to online safety, with a particular focus on protecting children and young people ages 13 to 17. Apple stated that through its financial support, the initiative aims to develop more effective technical solutions for reducing the risk of exposure to harmful content while maintaining user privacy, facilitate information sharing through opensource materials, and contribute to ongoing debates and discussions about online safety. These contributions, which include research briefs, issue spotting analyses, and technical descriptions authored by individual contributors with different backgrounds, are subject to peer review and will be made available for public comment.
- Apple was a member of, or provided general support to, numerous organizations dedicated to mitigating the risks of online harms and educating users about online safety, including the Tech Coalition, the Family Online Safety Institute ('FOSI'), INHOPE, Connect Safety, and the Digital Wellness Lab. Apple has supported these organizations for multiple years prior to this reporting period, including in their research activities.

Provide an explanation of alternative measures Apple Services took to detect known CSEA images on iMessage, FaceTime and iCloud

Apple stated that it had Communication Safety in place, which provided interventions when children received or attempted to send images that contain nudity and provided guidance and age-appropriate resources. Apple also stated that if a child was under 13 years of age, Communication Safety required that the Screen Time passcode for the Family Sharing group be entered on the child's device in order to view or send sensitive content, and this feature was

available on services including iMessage, AirDrop, FaceTime video messages, Contact Poster and the Photos picker.

Apple also stated that its Sensitive Content Warning feature provided similar protections for all Apple accounts by alerting users about nude images and videos when they are received.

Apple also stated that it had other online safety features including user-reporting mechanisms for unlawful or harmful content and 'Other illustrative examples of Apple's online safety features included but were not limited to user reporting mechanisms for unlawful or harmful content' and a 'Learn More about Online Safety' button in the settings for iMessage, FaceTime, iCloud and iCloud email for users in Australia.

Apple stated that it provided awareness material for parents and children including on Siri, Spotlight, and Safari Search which pointed to methods for reporting CSEA, and Apple's Personal Safety User guide and Child Safety Resources contained strategies, tools and how-to topics for dealing with online harms. Similarly, Apple retail locations hosted 'Your Kids and Their Devices' sessions for parents.

Was work undertaken to research, develop, source or implement tools to detect known CSEA videos on iMessage, FaceTime and iCloud?

Apple responded 'yes' to this question.

Provide details and specify the aims of the research and/or development conducted to detect known CSEA videos on iMessage, FaceTime and iCloud

Apple referred to the response it provided relating to the aims of the research and/or development conducted to detect known CSEA images on iMessage, FaceTime and iCloud, with the exception of Digital Wellness Lab. Apple stated that it had taken the same approach to CSEA videos as it had with CSEA images.

Provide an explanation of alternative measures Apple Services took to detect known CSEA videos on iMessage, FaceTime and iCloud

Apple referred to the response it provided relating to alternative measures it took to detect known CSEA images on iMessage, FaceTime and iCloud, with the exception of its industry collaboration with Digital Wellness Lab, which is not specifically directed to the issue of detecting and scanning for known CSEA material. Apple stated that it had taken the same approach to CSEA videos as it had with CSEA images.

Was any work undertaken to research, develop, source or implement tools to detect new CSEA material on iCloud and iCloud email?

Apple responded 'yes' to this question.

Provide details and specify the aims of the research and/or development conducted to detect new CSEA material on iCloud and iCloud email

Apple stated that it researched, designed and launched several enhancements to Communication Safety during the report period including:

- In iOS 18, iPadOS 18, and macOS Sequoia, which were released in September 2024, Communication Safety required that the Screen Time passcode for the Family Sharing group be entered on a child's device in order to view or send a sensitive image. Communication Safety would not proactively notify the family organizer that a child attempted to view or send a sensitive image, but it now requires that the family organizer enter the Screen Time passcode on the child's device if the content is to be viewed. This feature was developed with the aim of keeping children safe from inappropriate content and providing tools to parents to help keep their children safe from harmful content, including CSEA material.
- In December 2024, Apple enhanced Communication Safety in iMessage by allowing users in Australia to report inappropriate nudity flagged by Communication Safety directly to Apple. This feature was also developed and designed with the aim of keeping children, and all users, safe from inappropriate content and providing those users with the tools to report inappropriate nudity, potentially including new CSEA material, to Apple.
- Apple undertook research to assess the value and efficacy of Communication Safety.

Provide an explanation of alternative measures Apple Services took to detect new CSEA material on iCloud and iCloud email

Apple referred to the response it provided relating to alternative measures it took to detect known CSEA images on iMessage, FaceTime and iCloud, with the exception of Digital Wellness Lab.

Was any work undertaken to research, develop, source or implement tools to detect the livestreaming of CSEA on FaceTime?

Apple responded 'yes' to this question.

Provide details and specify the aims of the research and/or development conducted to detect the livestreaming of CSEA on FaceTime?

Apple stated that it made Communication Safety and Sensitive Content Warning available on FaceTime video messages to avoid the potential misuse of FaceTime by bad actors sending unwanted nude videos.

Apple stated that research was undertaken regularly to expand and improve the scope of detection for use in Apple Services.

Apple referred to the response it provided relating to alternative measures it took to detect known CSEA images on iMessage, FaceTime and iCloud.

Provide an explanation of alternative measures taken to detect CSEA livestreams on FaceTime, and any other relevant context

Apple referred to measures it took to detect known CSEA images on iMessage, FaceTime and iCloud and its response to the question 'Provide details and specify the aims of the research and/or development conducted to detect the livestreaming of CSEA on FaceTime?'

Apple stated that FaceTime was not designed for large livestreaming experiences, but rather as a secure and private way to make video calls. Apple stated that FaceTime creates a private, end-to-end encrypted connection between devices for real-time communication and that it requires explicit permissions and connections – that is, a recipient must actively accept your call request, each participant sees exactly who was in the call, and users maintain control by being able to decline calls or block contacts through their device settings. Apple stated that FaceTime calls are also limited by design to closed groups with no more than 32 participants.

Given that up to 32 participants at any one time are able to join a facetime call, eSafety notes that there exists the possibility that 32 participants could choose to engage in or view CSEA on a single call.

Proactively detecting CSEA activity – grooming

Provide an explanation of alternative measures Apple Services took to detect grooming on iCloud email, iMessage and FaceTime

Apple stated that the Communication Safety feature was designed to help stop the grooming of children by intervening and offering helpful resources to children if they receive or attempt to send images or videos containing nudity. If the child was under 13, Communication Safety required that the Screen Time passcode for the Family Sharing group be entered on the child's device in order to view or send sensitive content.

In addition, Apple stated that it introduced a new feature for iMessage in Australia that allowed children and adults to report nude images and video being sent to them directly to Apple, which could then appropriately report the issue to NCMEC and law enforcement agencies.

Apple stated that with Screen Time, parents could set limits for who the child could communicate with during screen time and downtime, and that these limits applied to phone calls, FaceTime, Messages and iCloud contacts. A parent could request permission to manage a child's iCloud contacts or limit the child's communication to certain contacts.

Apple also referred to its response to 'Provide an explanation of alternative measures Apple Services took to detect known CSEA images on iMessage, FaceTime and iCloud'.

Proactively detecting CSEA activity – sexual extortion

Provide an explanation of alternative measures Apple Services took to detect sexual extortion of adults and children on iCloud email, iMessage and FaceTime

Apple stated that its systems designed to detect fraud and spam also naturally helped identify sexual extortion attempts because these schemes often shared common elements such as suspicious payment patterns, mass messaging behaviour and unusual account activity. Apple stated that when it blocked accounts for fraud or spam violations, Apple was often simultaneously helping to stop sexual extortion operations that rely on these same deceptive tactics and communication methods.

Apple also stated that Apple users could block phone numbers, contacts, and emails, and that Apple's investment in grooming detection could help it to identify sexual extortion patterns as well.

Apple also referred to its response to 'Provide an explanation of alternative measures Apple Services took to detect known CSEA images on iMessage, FaceTime and iCloud'.

Blocking URLs to known CSEA material

Was any work undertaken to research, develop, source or implement tools to block URLs linking to known CSEA on iMessage, FaceTime, iCloud and iCloud email?

Apple responded 'no' to this question.

Provide an explanation of alternative measures Apple Services took to prevent URLs linking to known CSEA materials on iMessage, FaceTime, iCloud and iCloud email

Apple stated that Screen Time and Restrictions could be applied to web content, such as to block adult content and specific websites across all browsers in a child's phone. There was also an option for parents to block all websites, except for a specific set that they choose to allow for their children. This tool helped parents play a direct role in keeping their kids safe online, while allowing children to have access to valuable educational and entertainment resources. Apple stated that since iOS 17, adult websites have been blocked on child devices and parents can choose to allow or block additional websites.

Apple also referred to its response to 'Provide an explanation of alternative measures Apple Services took to detect known CSEA images on iMessage, FaceTime and iCloud'.

In-service user reporting

Could users report instances of CSEA to Apple in-service (that is, without being required to locate a separate webform or email address on another website)?

Service	Users could report instances of CSEA to Apple in-service
iCloud email	No
iCloud	No
iMessage	Yes[126]
FaceTime	No

Was any work undertaken with the goal of implementing in-service user reporting options for users to report CSEA material on iCloud email, iCloud and FaceTime?

Apple responded 'yes' to this question.

Provide details and specify the aims of this work and/or development conducted

Apple stated that it had implemented an enhancement to its Communication Safety feature that allowed children to report nude photos and video being sent to them in iMessage directly to Apple. The report was reviewed by Apple, which could then take action on an account — such as disabling that user's ability to send messages over iMessage — and also report the issue to law enforcement. Reporting to Apple was included in iOS 18.2, iPadOS 18.2, macOS 15.2 and visionOS 2.2 in Australia. Apple stated that because it believed this was an important feature for all users, Apple will be rolling it out globally in the future.

Apple stated that it developed a Content Reporting Portal (https://contentreports.apple.com/), which allowed users to submit reports about iCloud content. Users could access this portal directly or by clicking on a report button available in public files within iCloud.com Photos and Drive.

Apple's Communication Safety

Provide the proportion of Apple accounts used by children under the age of 13 in Australia that proceeded to view or send a sensitive image despite a Screen Time passcode being entered on the child's device[127]

Apple did not provide information that answered this question.

---

[126] Apple stated that users were able to report nude photos and videos received on iMessage directly to Apple through the Communication Safety feature. eSafety notes that this user reporting option required Communication Safety to have been enabled.

[127] Apple, 2023, 'Expanded Protections for Children', n.d., accessed 18 June 2024, URL: https://www.apple.com/child-safety/

Apple stated that its analytical collection was dependent on users opting in to providing data to Apple. This was done to preserve and protect privacy, which Apple considers an important dimension to online safety.

Apple stated that if a user opted in to 'Improve Communication Safety', this data was used to improve and understand the effectiveness of the Screen Time feature. This information was aggregated and subject to Apple's data collection process, but also required enough devices to share data in order to perform any meaningful analysis. eSafety understands this means that few users have opted in to 'Improve Communication Safety' and that as a result, Apple was unable to measure whether the Screen Time passcode was an effective safety measure in preventing CSEA and sexual extortion during the report period.

Apple stated that it was actively exploring methods to gather more meaningful metrics that will help Apple improve the feature which protect user privacy. This exploration included enhancing Apple's analytics collection implementation later in 2025 to better capture the following metrics:

- For all devices with Communications Safety enabled, the percentage of images/videos received in iMessage flagged as containing nudity
- For all devices with Communications Safety enabled, the percentage of images/videos flagged as containing nudity and blurred, that are tapped on by the user
- The percentage of user taps on options presented within the Communication Safety intervention screens ('Now Now', 'Ways to Get Help', 'I'm Sure', 'Don't View', 'Message Someone', 'View')
- For those who tapped 'View', the percentage of instances where a user is asked for the Screen Time PIN and the Screen Time PIN was entered successfully.

Apple stated that user reporting complemented Communications Safety by creating a feedback loop for product improvement by enabling Apple to understand systematic biases, adjust detection thresholds, and prioritise improvements based on usage patterns.

Provide the proportion of users aged 13 to 17 in Australia that had Communication Safety enabled

Apple did not provide information that answered this question.

Apple stated that in order to protect the privacy of its users, its systems did not allow Apple to determine the ages of users that had Communication Safety enabled.

Provide the proportion of users aged 13 to 17 in Australia, with Communication Safety enabled, that proceeded to view or send a sensitive image despite a Communication Safety warning message

Apple did not provide information that answered this question.

Apple stated that in order to protect the privacy of its users, its systems did not allow Apple to determine the ages of users that had Communication Safety enabled.

Apple stated that it was working on adding new ways to access data for Communication Safety as discussed above.

Specify the proportion of adult accounts[128] in Australia that enabled Communication Safety

Apple did not provide information that answered this question.

Apple stated that in order to protect the privacy of its users, its systems did not allow Apple to determine the ages of users that had Communication Safety enabled.

Specify all research undertaken to expand or improve Communication Safety on the Apple Services

Apple stated that the Communication Safety's machine learning-based classifier was being further trained and improved over time as new insights in model behavior and training data was acquired through external sources. The model was updated through software updates.

Apple stated that it conducted studies with users who have enabled Communication Safety and received and sent images and videos that had undergone interventions. These studies included direct user feedback when content was incorrectly flagged during normal usage, along with structured interviews with users who have experienced Communication Safety's intervention on the media, enabling targeted improvements to both the underlying machine learning models and the intervention UI/UX.

Apple stated that user reporting complemented Communications Safety by creating a feedback loop for product improvement by enabling Apple to understand systematic biases, adjust detection thresholds, and prioritise improvements based on usage patterns.

Apple also referred to the response it provided in relation to the aims of the research and/or development conducted to detect known CSEA images on iMessage, FaceTime and iCloud.

[128] Accounts held by users 18 and over

Resourcing to ensure safety

Languages that Provider's human moderators (employees and contractors) operated across

Languages covered by human moderators (employees):		
Chinese (Mandarin and Simplified Chinese)	Croatian	English
French	Spanish	

Languages covered by human moderators (contractors):		
Chinese (Mandarin and Simplified Chinese	English	Hindi

Discord

Key insights:

- Discord was not automatically notified when volunteer moderators removed users from servers for CSEA violations. Similarly, volunteer moderators were not notified by Discord when Discord banned users for CSEA. If a volunteer moderator removed a user from a server but a broader action wasn't taken by Discord, this could have allowed violative users to remain on the service to harm again in other servers.
- Discord did not have any in-service reporting options for livestreams or videocalls.

Compared with previous response[129]

In our 2023 transparency report, eSafety reported on Discord's processes its volunteer administrators/moderators follow, and the processes Discord has in place to monitor their conduct and uphold moderation standards.

As of this reporting period, Discord still did not have in place a policy for its volunteer administrators/moderators, there was still no way for users to make reports about volunteer administrators/moderators in instances where they had failed to effectively moderate relating to CSEA, and volunteer administrators/moderators were still not automatically notified when Discord's Trust and Safety team removed a user from a server for a violation relating to CSEA.

Additionally, since 2023, Discord has not put in place any in-service reporting options for livestreams or videocalls. This is particularly concerning, given Discord also implemented end-to-end encryption on its voice and video communications as of 17 September 2024.

Proactive detection of CSEA

Did Discord undertake any work to research, develop, source or implement tools to detect known CSEA videos on Discord?

Discord responded 'yes' to this question.

[129] A non-periodic notice was given to Discord in 2023. Note that the questions asked and the reporting timeframes differed between notice rounds. These comparisons serve as a guide to how service responses may have changed since the last notice they received. Future comparisons will be observed through the following three periodic notice responses.

Specify the aims of the research and/or development conducted to detect known CSEA videos on Discord

Discord stated that its safety team continuously evaluated the impact and effectiveness of new technologies, including hash-matching capabilities and novel CSEA detection for video uploads. Discord stated that it determined that the quality of databases and hash-matching abilities remains a crucial challenge. Discord stated that while video material can be directly uploaded by users, video was often shared as links or embedded videos from other platforms and that Discord had no visibility of these videos.

Discord stated that it implemented Google's Content Safety API into its technology stack and while this was not a hash-matching technology, it supported Discord's efforts to proactively identify video files sent on Discord that may contain known and new CSEA material or other content sexualising children. Discord used Google's Content Safety API to help classify and prioritise content for human review – the higher the priority given by the classifier, the more likely the media file contains abusive material.

Provide an explanation of alternative measures Discord Inc. took to detect known CSEA videos on Discord

Discord referred to its response specifying the aims of the research and/or development conducted to detect known CSEA videos.

Did Discord Inc. undertake any work to research, develop, source or implement tools to detect the livestreaming of CSEA?

Discord responded 'no' to this question.

Provide details, and specify the aims of the research and/or development conducted.

Discord's safety teams may have conducted contextual reviews of servers or direct messages based on information from organisations in Discord's Safety Reporting Network, or other trusted partners.

Proactively detecting CSEA activity – grooming

Languages that language analysis technology used to detect likely terms, abbreviations, codes, and hashtags indicating grooming operated in

Discord stated that as an experimental feature, the tool was limited to users whose language was set to English.

Did Discord undertake any work during the reporting interval to research, develop or source tools to detect grooming on Discord?

Discord responded 'no' to this question.

Provide an explanation of alternative measures Discord took to detect grooming on Discord, and any other relevant context

Discord stated that its alternative measures to detect grooming included investment in specialised teams focused on child safety and information sharing partnerships. Discord stated that its minor safety experts, were trained to proactively identify and appropriately action content and behaviours that posed risks to teens on Discord and that violated its policies prohibiting CSAM, child sexualization, and inappropriate sexual conduct with children.

Discord also stated that it participated in Lantern, and operated its Safety Reporting Network to collaborate with trusted reporters worldwide, empowering non-governmental organisations and government agencies with direct access to escalate content concerns.

Proactively detecting CSEA activity – sexual extortion

Languages that language analysis technology used to detect likely terms, abbreviations, codes, and hashtags indicating sexual extortion operated in

Discord stated that its language analysis tool to detect sexual extortion was language agnostic.

Did Discord undertake any work to research, develop, source or implement tools to detect sexual extortion on the service?

Discord responded ‘yes’ to this question.

Specify the aims of the research and/or development conducted to detect sexual extortion on the service

Discord referred to its response to questions shown in table 14. Discord also operated a voluntary Safety Reporting Network, allowing it to collaborate with various organisations to identify and report violations of Discord’s Community Guidelines. Members of the Network had access to prioritised reporting channels. Once Discord’s Safety team was made aware of policy violations, it could take actions including removing content, banning users, shutting down servers and engaging with authorities when appropriate. Discord stated that it partnered with many organisations through the Network including Australian partners such as Kids Helpline and the eSafety Commissioner as a Trusted Flagger.

Discord stated that the Safety Alerts in Chat feature were not active for all teen users during the entire report period. Discord paused the feature in the second half of 2024 (that is, during the report period) to evaluate its effectiveness. Discord stated that it intends to further refine the underlying model for future experiments.

The Safety Alerts in Chats test ran from April 2024 through July 2024.

Provide an explanation of alternative measures Discord took to detect sexual extortion of adults and children

Discord stated that its alternative measures to detect sexual extortion included investment in specialised teams focused on child safety and information sharing partnerships. Discord stated that its minor safety experts, were trained to proactively identify and appropriately action content and behaviours that posed risks to teens on Discord and that violated its policies prohibiting CSAM, child sexualization, and inappropriate sexual conduct with children.

Discord also stated that it participated in Lantern, and operated its Safety Reporting Network to collaborate with trusted reporters worldwide, empowering non-governmental organisations and government agencies with direct access to escalate content concerns.

Blocking links to known CSEA material

Did Discord undertake any work to research, develop, source or implement tools to block URLs linking to known CSEA material?

Discord responded ‘yes’ to this question.

Specify the aims of the research and/or development conducted to block URLs linking o known CSEA material

Discord stated that it partnered with IWF and had been evaluating their URL hash-matching database for both automatic blocking and in-service user reported content.

Provide an explanation of alternative measures Discord took to block URLs linking to known CSEA, and any other relevant context

Discord stated that its alternative measures to block URLs linking to known CSEA included investment in specialised teams focused on child safety and information sharing partnerships. Discord stated that its minor safety experts, were trained to proactively identify and appropriately action content and behaviours that posed risks to teens on Discord and that violated its policies prohibiting CSAM, child sexualization, and inappropriate sexual conduct with children.

Discord also stated that it participated in Lantern and operated its Safety Reporting Network to collaborate with trusted reporters worldwide, empowering non-governmental organisations and government agencies with direct access to escalate content concerns.

In-service user reporting

Could users report instances of CSEA during live video on Discord (that is, without being required to locate a separate webform or email address)?

Go Live	No
Video calls	No

Did Discord undertake any work with the goal of implementing in-service reporting options for users to report CSEA activity and/or material?

Discord responded 'yes' to this question.

If yes, specify the aims of the research and/or development conducted. Include whether this work was conducted on parts of a service (for example livestreaming on a public server)

Discord stated that its leadership was evaluating various methods of moderation for livestream and/or voice communications in Discord voice channels and calls that maintain Discord's commitment to end-to-end encryption.

Explain how Discord's user reporting options are clear and readily identifiable and provide any other relevant context

Discord stated that its voice and video communications were designed for small groups and were end-to-end encrypted. Discord was evaluating moderation tools for livestreams and/or voice communications. Organisations participating in Discord's Safety Reporting Network could report activity across the service to its safety teams – including activity from livestreams – that could lead to the investigation and removal of those spaces and users.

eSafety notes that up to 50 people can share their screen or video in a voice channel.

Information sharing between Volunteer Moderators and Trust and Safety staff relating to CSEA

Did Discord have a standards policy, or similar, outlining the responsibilities and expectations of the Discord volunteer administrator/moderator role?

Discord responded 'no' to this question.

Discord stated that Volunteer administrators and moderators were held to the same standard under the Terms of Service and Community Guidelines, and if they were found to be in violation, appropriate action was taken just as would occur for any other user. Community moderation allowed individual community administrators and moderators to establish rules on

top of Discord's Community Guidelines. If Discord learned of and investigated communities dedicated to content or activity that violated its Community Guidelines, Discord removed them from the service.

Discord provided a variety of tools and resources to server owners and moderators – as well as third-party applications created by third-party developers – that supported the development and enforcement of a community's rules. Features such as 'AutoMod', which allowed communities to implement community-specific word filters, helped better set the tone for their individual communities.

Server owners and moderators removed content at their own discretion and based on their community norms and values. This type of moderation was in addition to, and not instead of, other moderation conducted on the platform. Should a user have been removed by a private server moderator for violating server rules they could join other topic specific servers or start their own.

Was Discord's Trust and Safety team automatically notified when a volunteer administrator/moderator removed a user from a server for a violation relating to CSEA activity and/or material?

Discord responded 'no' to this question. Discord stated that moderators could implement their own rules and guidelines that went beyond the Discord Community Guidelines, so moderator actions on Discord were not automatically reported to Discord's Trust and Safety team.

Instead, moderators, like all users on Discord, were encouraged to report content that violated Discord's Community Guidelines.

Discord relied on a combination of proactive monitoring signals and user reports to identify activity that might have violated its Terms of Service, Community Guidelines, or policies generally. Server-specific moderation activity covered a wide variety of activity on a server, where violations of server rules may have had no relationship to a breach of broader Community Guidelines or Terms of Service. For example, server-specific rules may have governed content relevancy, media spoilers, language conventions, and other specific rules to which Discord did not have a reasonable need to intervene or adjudicate. While more serious harms that occur in-server may have also breached server-specific rules, involvement from Discord would have been on the basis of activity breaching Community Guidelines or Terms of Service.

Discord invested in features and resources that users could use to improve their community moderation efforts. If Discord learned of activity in communities that violated Discord's policies, Discord would have taken appropriate action which may include banning all members of a community and shutting down the server.

Were volunteer administrators/moderators automatically notified when Discord's Trust and Safety team removed a user from a server for a violation relating to CSEA activity and/or material?

Discord responded 'no' to this question. Discord stated that where a user violated Discord's policy related to CSEA, it was Discord that took action by banning the user from the service, making relevant reports to NCMEC or other appropriate authorities, and processing their account for deletion. If the purpose of a server was determined to be centred around CSEA material, Discord would action the entire server. Under other circumstances, when an entire server was not actioned, Discord safety teams would delete specific channels and all content contained in them. The administrator and/or all users in the server may have also been issued a server warning, reminded of Discord's Community Guidelines, and that further violations of the policy within their community could result in the termination of the community, the administrator's account, and the users.

If no, was work undertaken to develop a means of automatically notifying volunteer administrators/moderators when Discord's Trust and Safety removes users for CSEA-related reasons?

Discord responded 'no' to this question.

Were users able to make reports about volunteer administrators/moderators in instances where they had failed to effectively moderate relating to CSEA activity and/or material?

Discord responded 'no' to this question. Discord stated that users could submit reports regarding activity on Discord either through in-app reporting where available or through its standard report form on Discord.com.

Limiting the availability of AI-generated CSEA

How did Discord ensure that it's generative AI features would not produce CSEA material?

Discord stated that when it received a report of CSEA generated through generative AI apps on Discord, Discord's Developer Compliance team then reviewed the report. If CSEA was identified, the content was removed and reported to NCMEC (consistent with Discord's policies). If the content was attributed to a user's prompt, the user's account would also have been actioned. Discord's Developer Compliance team also assessed whether reports suggest the app's purpose was to enable the creation of CSEA, or whether the developer was not doing enough to prevent this abusive use (for example, based on relative volume of CSAM reports). An app would have been removed from Discord's developer platform upon such review.

Explain why Discord did not take measures to ensure that new CSEA material was not included in the data sets used to train generative AI models available on Discord?

Discord stated that its Developer Policy specified that developers must ensure that their apps do not, and are not used to, violate Discord's Community Guidelines or any other terms and policies. Developers and their apps were required to comply with the Developer Policy and Developer Terms of Service.

Resourcing to ensure safety

Languages that Discord's human moderators (employees and contractors) operated across

Discord's employees and contractors operated across the following languages:		
Brazilian Portuguese	Greek	Russian
Bulgarian	Hindi	Simplified Chinese
Croatian	Hungarian	Spanish
Czech	Italian	Swedish
Danish	Japanese	Thai
Dutch	Korean	Traditional Chinese
English	Lithuanian	Turkish
Finnish	Norwegian	Ukrainian
French	Polish	
German	Romanian	

Google

Key insights:

- Google did not have any in-service reporting options for Google Messages or Gmail.

Compared with previous response[130]

In our 2023 transparency report, eSafety also reported on the lack of in-service reporting options for Google Messages and Gmail.

Proactive detection of CSEA

Did Google undertake any work to research, develop, source or implement tools to detect known CSEA images on Messages?

Google responded 'no' to this question.

What alternative measures did Google take to detect known CSEA on Messages?

Google stated that it had strict policies that prohibited the use of its services to share or distribute CSEA, and made efforts to detect, remove and report CSEA material across all of its services.

Google stated that Messages included, in some instances, end-to-end-encrypted communications where detection was more technically challenging and subject to heightened privacy and security considerations. Google stated that its tools to intervene in encrypted spaces included, among other things, responding to user reports and using signals. Google stated that in addition to its use of hash-matching technology to detect known CSAM (on relevant services), it also had the following protections in place across all of its services:

- provided information to help users report child sexual abuse to the relevant authorities by geographical region (which for Australia, includes the details of the eSafety Commissioner). Anyone could report abuse to Google by completing its child endangerment webform that was available via the Google Help Centre (this was in addition to the ability to report abuse in product, where available).
- invested significantly in developing strong Trust & Safety Teams with operators who provided cover 24 hours a day to quickly respond in the case of any child safety incident, and where

[130] A non-periodic notice was given to Google in 2023. Note that the questions asked and the reporting timeframes differed between notice rounds. These comparisons serve as a guide to how service responses may have changed since the last notice they received. Future comparisons will be observed through the following three periodic notice responses.

there may be an imminent harm to a child, Google's specialist team escalated the report to NCMEC's prioritisation queue.

Google stated that it was a member of the Technology Coalition which, as part of Project Protect – a cross-industry initiative to combat CSAM through investment, research and information sharing – announced research grants for institutions working on actionable research including hot to better detect online grooming. Google stated that it would be monitoring the outputs from this initiative and considering how they could be applied to Google services in the future.

Did Google undertake any work to research, develop, source or implement tools to detect known CSEA videos on Chat, Messages, Gmail or Gemini?

Google responded 'no' to this question.

What alternative measures did Google take to detect known child sexual exploitation and abuse videos on Chat, Messages, Gmail and Gemini?

Google stated that detection of CSAM videos can be more challenging than image processing and detection in certain services, and while Google deployed this on Drive and YouTube during the report period, it stated that it continues to evaluate its deployment in products such as Chat and Gmail.

Gmail and Chat employed a specific approach for handling larger files. Direct attachments were limited to 25MB in Gmail and 200MB in Chat. Files exceeding these limits were automatically stored on Google Drive, with only a shareable link provided within Gmail or Chat. Google Drive utilised CSAI Match to scan videos before they could be shared via these links, preventing the distribution of known CSAM, even when shared indirectly through Drive links in Gmail and Chat.

While Google did not screen for known CSAM videos on Chat and Gmail, it did screen for known CSEA images using hash-matching technologies. If known CSAM images were detected in association with an account, the account may have been disabled, and facts and circumstances of the CSAM offence reported to NCMEC, who may have then forwarded those reports to relevant law enforcement agencies globally. Google stated that in many instances, a user may have been sharing or distributing both known CSAM images and known CSEA videos in which case they may still be caught by this process. Similarly, if CSAM was detected on another Google service associated with an account, the Google account would be disabled and a report to NCMEC would be made.

Google stated that Gemini did not allow users to upload or create any videos.

Did Google undertake any work to research, develop, source or implement tools to detect new CSEA on Meet, Chat, Messages, Gmail or Gemini?

Google responded 'yes' to this question.

Provide details and specify the aims of the research and/or development conducted to detect new CSEA on Meet, Chat, Messages, Gmail or Gemini?

Google stated that it was researching the use of machine learning classifiers on these Google services. The objective was to broaden machine learning classifiers' usage across Google services, where feasible.

What alternative measures did Google take to detect new CSEA videos on Meet, Chat, Messages, Gmail or Gemini?

Google stated that if known CSAM content was detected in association with a Google account, the account may have been disabled, and the facts and circumstances of the CSAM offence reported to NCMEC, who then forwarded those reports to relevant law enforcement agencies globally. This meant that if an account was used for both new and known CSAM, it may have still be detected via these processes.

Similarly, if new CSAM content was detected on other Google services that were used by that user in connection with that same account, their Google account could be disabled and a report to NCMEC would be made.

Finally, where new CSAM was detected on any of Google's services, it was hashed and added to Google's internal repository, and to the extent hash-matching was performed on these products, could then identify duplicates and near-duplicates in its other services to take action accordingly, including reporting, preserving, and deleting the content and taking account level enforcement actions where appropriate.

Google stated that Gemini did not allow users to upload or create any videos.

Languages of tools used to detect CSEA livestreaming on YouTube:

Afrikaans	Gujarati	Piedmontese
Albanian	Haitian	Polish
Arabic	Hebrew	Portuguese
Aragonese	Hindi	Punjabi
Armenian	Hungarian	Romanian
Asturian	Icelandic	Russian
Azerbaijani	Ido	Scots
Bashkir	Indonesian	Serbian

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

Did Google undertake any work to research, develop, source or implement tools to detect the livestreaming of CSEA on Meet?

Google responded ‘no’ to this question.

What alternative measures did Google take to detect CSEA livestreams

Google stated that content on Google Meet was always reviewed when it was reported. Along with reporting a meeting participant, users had the option to upload a short video clip to substantiate the claim of abuse. The meeting report was sent for human review and appropriate actions were taken on the user(s) found to violate terms of service for Google Meet.

Proactively detecting CSEA activity – sexual extortion and grooming

Did Google undertake any work to research, develop, source or implement tools to detect sexual extortion of adults on Meet, Chat or Messages?

Google responded 'no' to this question.

Did Google undertake any work to research, develop, source or implement tools to detect grooming on Meet, Chat or Messages?

Google responded 'no' to this question.

Languages of tools used to detect sexual extortion and grooming on YouTube:

Afrikaans	Gujarati	Piedmontese
Albanian	Haitian	Polish
Arabic	Hebrew	Portuguese
Aragonese	Hindi	Punjabi
Armenian	Hungarian	Romanian
Asturian	Icelandic	Russian
Azerbaijani	Ido	Scots
Bashkir	Indonesian	Serbian
Basque	Irish	Serbo-Croatian
Bavarian	Italian	Sicilian
Belarusian	Japanese	Slovak
Bengali	Javanese	Slovenian
Bishnupriya Manipuri	Kannada	South Azerbaijani
Bosnian	Kazakh	Spanish
Breton	Kirghiz	Sundanese
Bulgarian	Korean	Swahili
Burmese	Latin	Swedish
Catalan	Latvian	Tagalog
Cebuano	Lithuanian	Tajik
Chechen	Lombard	Tamil
Chinese (Simplified)	Low Saxon	Tatar
Chinese (Traditional)	Luxembourgish	Telugu
Chuvash	Macedonian	Thai
Croatian	Malagasy	Turkish
Czech	Malay	Ukrainian
Danish	Malayalam	Urdu

Dutch	Marathi	Uzbek
English	Minangkabau	Vietnamese
Estonian	Mongolian	Volapük
Finnish	Nepali	Waray-Waray
French	Newar	Welsh
Galician	Norwegian (Bokmal)	West Frisian
Georgian	Norwegian (Nynorsk)	Western Punjabi
German	Occitan	Yoruba
Greek	Persian (Farsi)	

What alternative measures did Google undertake any work to research, develop, source or implement tools to detect sexual extortion of adults and children and grooming on Meet, Chat or Messages?

Google stated that it aimed to prevent abuse from happening by ensuring that its products are safe while remaining aware of heightened privacy and legal considerations in using language-based analysis tools for scanning of private electronic communications (for instance, Chat) that may not apply to services such as YouTube.

Google stated that its protective measures in place to protect minors from sexual extortion and grooming for each of Meet, Chat, and Messages included:

- a user could only be contacted by another user of those services if that user knows or has been provided with that user's email address / phone number
- a user could at any time block another user from contacting them via the service. Google published information about how a parent or child could block a user on each of these services
- as part of Google's work to make its products safe for children to use, Google provided information to help users report child sexual abuse material to the relevant authorities by geographical region. Parents or the child could also report abuse, which included child grooming, either in product (if available), or by completing Google's child endangerment webform that was available via the Google Help Centre. These reports were handled by its team of child safety experts and specialists and may involve reporting to NCMEC and appropriate account-level enforcement. Additionally, to the extent evidence of grooming, sexual extortion, or child trafficking was detected through Google's proactive CSAM workflow, Google's child safety team took action including possible disabling of the account as well as reporting to NCMEC or to law enforcement.

Google stated that while a minor under 13 years old was not permitted to have a Google Account, Family Link allowed parents to create and supervise an account on behalf of their child. This included signing in or out of a child's account, blocking contacts for a child on Meet

or Chat, choosing apps that a child can see and/or approving download of apps on Google Play (which may include preventing a child from downloading apps that allow for messaging), and managing content settings on services such as Search, YouTube and Play. Google stated that parental controls for Google Accounts continue to be available for minors aged 13 - 17 years. Also, and in addition to the use of language analysis technology on YouTube to detect sexual extortion and grooming, other protective measures in place to protect minors from 'sexual extortion' and 'grooming' on YouTube included:

- The default upload settings for users aged 13 to 17 was set at 'private'. Private videos and playlists can only be seen by the user and whomever they choose. A user's private videos did not appear in the Videos tab of a user's channel homepage. They also did not show up in YouTube's search results
- Content that did not violate Google's policies but features children may have some features disabled at both the channel and video level. These features may include: comments, live chat, livestreaming, video recommendations (where and when a video is recommended) and community posts
- With a supervised account, parents could select a content setting that limits the videos and music children under 13 can find and play, and the features that could be used (including comments, handles and live-chats).

eSafety understands from Google's response that it did not take alternative steps to proactively detect sexual extortion on Meet, Chat and Message. However, Google stated that it, took steps to prevent sexual extortion of adults and children on its services.

Blocking URLs to known CSEA material

Did Google undertake any work to research, develop, source or implement tools to block URLs linking to known CSEA material?

Google responded 'yes' to this question.

Provide details and specify the aims of the research and/or development conducted to block URLs linking to known CSEA material

Google stated that it maintained its own list of URLs that were linked to known CSAM from Google's effort to identify CSAM on Search. Google stated that when it identified a URL linking to CSAM material, Google reported that URL to NCMEC, and then de-indexed that URL from surfacing on Search. Google also investigated URLs reported by third parties - such as specialist child safety organisations - that were suspected to lead to CSAM and may report and de-index these URLs.

What alternative measures did Google take to prevent URLs linking to known CSEA materials?

Google stated that when it identified URLs that contain CSAM, including reports received from priority flaggers such as NGOs and government agencies, Google reported that URL to NCMEC and de-indexed it from surfacing on Google Search.

Google stated that although it did not proactively scan for URLs on its services, it did hash CSAM content identified in those services and took action accordingly, including reporting, preserving, and deleting the content and taking account level enforcement actions where appropriate.

Google stated that due to the changeable nature of content that may appear at a URL, hash-matching was an effective way of detecting and preventing the re-circulation of the CSAM that may appear at these links across Google services. Google stated that it also had strict policies that prohibit the use of its services to share or distribute CSEA material (including the sharing of links to known CSAM) which included taking enforcement action on solicitation or advertisement of CSAM; therefore, if Google became aware of a URL link that leads or purports to lead to CSAM, it would report and remove it and take other enforcement actions as appropriate

Preventing the spread and creation of CSEA – recommender systems

Did Google use the following measures to test and update its YouTube recommender systems to contribute to the overall safety of the service and avoid the risk of amplifying unlawful and harmful content?

Testing	YouTube
Internal audits	No
External audits	No
Risk assessments	No
Impact assessments	Yes
a/b testing	No
Other	Third-party evaluators

How often were the tests to YouTube's recommender system performed to contribute to the overall safety of the service and avoid the risk of amplifying unlawful and harmful content?

Google stated that its engineering teams continually experimented and tested improvements to search results and recommendations. Google used these test results, along with user feedback and other signals, to update and improve recommendations over time.

YouTube also used third-party evaluators on a continuous basis to provide feedback on YouTube's search results, recommendations, and the relevance of certain videos. This feedback helped Google evaluate, train, and improve its systems, as well as the quality of recommendations on YouTube. Input from third-party evaluators was gathered and fed into reports that were generated weekly.

These evaluators were trained using the same public guidelines that Google uses for search results. These evaluators were different from Google's Community Guideline reviewers, who focus on enforcing Google's content policies and removing videos that violate Google's Community Guidelines.

In addition, YouTube used the violative view rate metric to measure the effectiveness of Google's enforcement of the YouTube Community Guidelines. The violative view rate, appeals and reinstatements metrics were calculated quarterly and published as part of Google's ongoing transparency reporting.

This metric was an estimate of the proportion of video views that violate Google's Community Guidelines in a given quarter (excluding spam). In order to calculate the violative view rate, Google took a sample of the views on YouTube and sent the sampled videos for review. Once Google received the decisions from reviewers about which videos in the sample are violative, it aggregated these decisions in order to arrive at its estimate. For Q3 2024, the number was 0.10 - 0.11%, which means that for every 10,000 views, between 10 and 11 were of content that violated Google's Community Guidelines.

Google also tracked the number of appeals submitted by creators in response to videos that were removed (an option available to any Creator on YouTube), as this helped Google gain a clearer understanding about the accuracy of its systems. For example, for the period July through September 2024, Google removed 28,159 videos that were uploaded in Australia for violating its Community Guidelines and received roughly 2,111 removal appeals. Upon review, Google reinstated 322 appeals. Google stated that these metrics were in addition to ongoing work to review and update the YouTube Community Guidelines to address new risks or harms as and when they emerge. To do this, Google worked with NGOs, academics, and relevant experts from all sides and different countries to inform this policy review. They help flag new concerns, or bring a deep understanding to complex topics that are prone to consistent change.

In-service user reporting

Could users in Australia report instances of CSEA activity and/or material on the following services in-service? (that is, without being required to locate a separate webform or email address on another website)?

Gmail	No
Messages	No
Gemini	Yes

Did Google undertake any work during the reporting interval with the goal of implementing in-service reporting options for users to report CSEA activity and/or material on Gmail or Messages?

Gmail	Yes
Messages	No

Specify the aims of the research and/or development conducted with the goal of implementing in-service reporting options for users to report CSEA activity and/or material on Gmail

Google stated that a reporting option was tested with Gmail users to determine the efficacy of including the in-service reporting link in addition to the central reporting form.

Provide an explanation as to how Google's user reporting features were clear and readily identifiable, and any other relevant context

Google stated that it encouraged its users first and foremost to report CSAM material or potential abuse, exploitation or trafficking to law enforcement, or NCMEC or its local counterparts in other countries. Google stated that as part of its work to make its products safe for children to use, Google provided useful information to help users report child sexual abuse material to the relevant authorities by geographical region. Where in-service reporting was not available, users could report if Google products were being used for child endangerment via a dedicated webform.

Could users in Australia make reports on YouTube without signing in?

Google responded 'no' to this question.

Did Google undertake any work during the reporting interval with the goal of implementing user reporting on YouTube in Australia for users who are not signed in?

Google responded 'no' to this question.

Resourcing to ensure safety

Languages covered by human moderators (employees):

English

Languages covered by human moderators (contractors):

Afrikaans	Hindi	Portuguese-BR
Amharic	Hungarian	Punjabi
Arabic	Igbo	Romanian
Azerbaijani	'Indian Languages'	Russian
Belarusian	Indonesian	Serbian
Bengali	Irish	Sinhalese
Bosnian	Italian	Slovenian
Bulgarian	Japanese	Somali
Burmese	Kazakh	Spanish
Cantonese	Khmer	Swahili
Croatian	Korean	Swedish
Czech	Kurdish	Tagalog
Danish	Laos	Tajik
Dutch	Latvian	Tamil
English	Lithuanian	Telugu
Estonian	Macedonian	Thai
Ethiopian-Amharic	Malay	Tigrinya
Ethiopian-Oromo	Malayalam	Turkish
Ethiopian-Tigriniya	Mandarin	Ukrainian
Filipino	Marathi	Urdu
Finnish	Norwegian	Uyghur
French	Oriya	Uzbek
German	Oromo	Vietnamese
Greek	Pashto	Yoruba
Gujarati	Persian	Zulu
Hausa	Polish	
Hebrew	Portuguese	

Google also stated that it had language agnostic reviewers. Google stated that agnostic reviews are primarily performed where no specific language proficiency was needed to conduct the review (for example, reviews of adult content), or in certain cases where Google couldn't identify the language.)

Meta

Key insights:

- Less than 1% of children whose accounts reflected that they were aged 13 to 17 or 13 to 15 on Facebook and Facebook Messenger had parental supervision enabled. 1% of those whose accounts reflected that they were 13 to 17 and 3.2% of those whose accounts reflected that they were 13 to 15 on Instagram had parental supervision enabled.

Compared with previous response[131]

In our 2022 transparency report, eSafety reported that Meta shared Facebook information with WhatsApp for the purpose of preventing unlawful or harmful material and activity, but noted that Instagram did not share information with WhatsApp for these purposes. It is therefore a positive improvement that Meta now reported sharing information from Instagram with WhatsApp when an account was banned or suspended for CSEA.

Proactive detection of CSEA

Languages of tools used to detect CSEA livestreaming on Facebook Live and Instagram Live

Languages covered by audio transcription:		
Arabic	Italian	Sinhala
Bengali	Japanese	Spanish
Burmese	Kannada	Tamil
English	Malay	Thai
French	Malayalam	Turkish
German	Marathi	Urdu
Hindi	Portuguese	Vietnamese
Indonesian	Russian	

[131] A non-periodic notice was given to Meta in 2022. Note that the questions asked and the reporting timeframes differed between notice rounds. These comparisons serve as a guide to how service responses may have changed since the last notice they received. Future comparisons will be observed through the following three periodic notice responses.

Languages covered by Internal proprietary tools:

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

Languages covered by Internal proprietary tools:

Afrikaans	Hungarian	Sindhi
Albanian	Icelandic	Sinhala
Amharic	Indonesian	Slovak
Arabic	Irish	Slovenian
Armenian	Italian	Somali
Assamese	Japanese	Spanish
Azerbaijani	Javanese	Sudanese
Basque	Kannada	Swahili
Belarusian	Kazakh	Swedish
Bengali	Khmer	Tamil
Bengali Romanised	Korean	Tamil Romanised
Bosnian	Kurdish (Kurmanji)	Telugu
Breton	Kyrgyz	Telugu Romanised
Bulgarian	Lao	Thai
Burmese	Latin	Turkish
Catalan	Latvian	Ukrainian
Chinese (Simplified)	Lithuanian	Urdu
Chinese (Traditional)	Macedonian	Urdu Romanised
Croatian	Malagasy	Uyghur
Czech	Malay	Uzbek

Danish	Malayalam	Vietnamese
Dutch	Marathi	Welsh
English	Mongolian	Wester Frisian
Esperanto	Nepali	Xhosa
Estonian	Norwegian	Yiddish
Filipino	Oriya	
Finnish	Oromo	
French	Pashto	
Galacian	Persian	
Georgian	Polish	
German	Portuguese	
Greek	Punjabi	
Gujarati	Romanian	
Hausa	Russian	
Hebrew	Sanskrit	
Hindi	Scottish Gaelic	
Hindi Romanised	Serbian	

Did Meta undertake any work to research, develop, source or implement tools to detect the livestreaming of child sexual exploitation and abuse on Facebook Messenger calls?

Meta responded ‘no’ to this question.

Provide an explanation of alternative measures Meta took to detect CSEA livestreams on Facebook Messenger calls

Meta stated that it differentiated between ‘livestreaming’ products which were 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 were designed to enable a user to have a private interpersonal end-to-end encrypted conversation with another user or a small group of users.

Meta stated that while it implemented a range of measures to detect the livestreaming of CSEA in its livestreaming products, in order to protect the privacy of its users and to comply with applicable law, including the U.S. Wiretap Act, Meta did not proactively monitor private calls on video calling products like Facebook Messenger. In addition, where end-to-end encryption was deployed, Meta was unable to access the underlying content of messages on these services, which meant proactive monitoring was not possible.

Use of language analysis tools

The proportion of all reports generated by automated tools used (which may or may not include language analysis tools), which were reviewed by human reviewers

Service	Proportion of all reports generated by automated tools which were reviewed by human moderators (approximate percentage)
Facebook	3%
Instagram (including Instagram Direct)	6%
Facebook Messenger	1%
Threads	20%

Meta stated that 'reports' referred to content flagged by Meta's systems as potentially violating its policies, which were then reviewed by human moderators (but not any content flagged by user reports). Meta stated that the vast majority of content proactively flagged by Meta's systems was removed automatically, leaving only a very small subset of such content to be reviewed by human moderators.

If human moderators did not review all reports generated by automated tools, provide an explanation of the threshold(s) for determining whether human review was needed

Meta stated that its automated tools generated signals that reflected the likelihood that material or activity may have violated one or more of Meta's policies. If there was a high likelihood of a violation, Meta stated that its systems would likely automatically take enforcement action in response to that violation. If Meta's systems did not make an automated enforcement decision, the relevant material or activity was enqueued for human review. Meta used the signals from its automated tools to create a prioritisation score, which enabled Meta to prioritise reports for human review based on severity, virality and likelihood of violation. Meta stated that there was no threshold or cut off score that determined whether or not a given report was subject to human review. Rather, the system operated in a dynamic manner with all reports enqueued in order of priority, and the order adjusted as new reports were enqueued. Meta stated that even if a report was not reviewed on day one, it remained in the queue and may have been reviewed on a subsequent day, depending on its prioritisation score relative to any new reports that were enqueued. Meta stated this ensured that the human review team was always reviewing reports of most concern at the relevant point in time.

Proactively detecting CSEA activity – grooming

The proportion of all reports generated by automated tools used (which may or may not include language analysis tools), which were reviewed by human reviewers and which were confirmed as violating Meta's policies around grooming

Service	Proportion of all reports generated by automated tools which were reviewed by human moderators (approximate percentage)
Facebook	1%
Instagram (including Instagram Direct)	1%
Facebook Messenger	1%
Threads	1%

Proactively detecting CSEA activity – sexual extortion

The proportion of all reports generated by automated tools used (which may or may not include language analysis tools), which were reviewed by human reviewers and which were confirmed as violating Meta's policies around sexual extortion

Service	Proportion of all reports generated by automated tools which were reviewed by human moderators (approximate percentage)
Facebook	1%
Instagram (including Instagram Direct)	1%
Facebook Messenger	1%
Threads	1%

Proactively detecting CSEA material and activity on end-to-end encrypted services

Did Meta undertake any work to investigate methods of detecting known CSEA on the end-to-end encrypted parts of Meta Services?

Meta responded 'no' to this question. Meta stated that as it is not technically possible to detect known CSEA on end-to-end encrypted parts of the service, Meta has invested in other measures to detect CSEA on end-to-end encrypted parts of the service, including the use of behavioural and account-level signals.

Preventing the spread and creation of CSEA – recommender systems

Did Meta use the following measures to test and update its recommender systems to contribute to the overall safety of the service and avoid the risk of amplifying unlawful and harmful content?

Testing	Facebook Feed recommender system	Instagram Reels and Feed recommender system	Threads 'For You' Feed recommender system
Internal audits	Yes	Yes	Yes
External audits	Yes	Yes	Yes
Impact assessments	Yes	Yes	Yes
a/b testing	Yes	Yes	Yes

How often were these tests of recommender systems performed?

Meta stated these tests were performed annually, at a minimum.

Sharing information to reduce CSEA (and sexual extortion of adults)

Regarding Take It Down and Stop NCII, specify whether Meta took and used all available hashes, or a subset, on Meta Services?

Initiative	Meta took and used all available hashes, or a subset
Take It Down	All available hashes
Stop NCII	All available hashes

Resourcing to ensure safety

Languages that Meta human moderators (employees) operated across:

Albanian	French (Sub-Saharan Africa)	Oromo
Amharic	Fula	Pashto (Afghanistan)
Arabic (Gulf)	Georgian	Persian
Arabic (Levant, Egypt, Iraq)	German	Polish
Arabic (Sudan)	Greek	Portuguese
Armenian	Gujarati	Punjabi
Assamese	Hausa	Romanian
Azerbaijani	Hebrew	Russian
Bambara	Hindi	Serbian

Belarusian	Hungarian	Sindhi (Pakistan)
Bemba	Igbo	Sinhala
Bengali	Indonesian	Somali
Bengali (India)	Italian	Spanish
Bosnian	Japanese	Swahili
Bulgarian	Kannada	Swedish
Burmese	Kazakh	Tamil
Cambodian	Kirundi	Telugu
Cantonese	Kituba	Thai
Chinese (Simplified)	Korean	Tigrinya
Chinese (Traditional)	Kurdish	Tok Pisin
Croatian	Latvian	Turkish
Czech & Slovak	Lingala	Ukrainian
Danish	Lithuanian	Urdu (India)
Dari (Afghanistan)	Malayalam	Urdu (Pakistan)
Dutch	Malaysian	Vietnamese
English	Marathi	Yoruba
Estonian	Mongolian	Zulu
Filipino	Nepali	
French	Norwegian	

Languages that Meta human moderators (contractors) operated across:

Afrikaans	Gujarati	Oriya
Albanian	Hausa	Oromo
Amharic	Hebrew	Pashto, Pushto
Arabic	Hindi	Persian
Armenian	Hungarian	Polish
Assamese	Indonesian	Portuguese
Azerbaijani	Italian	Punjabi
Bengali	Japanese	Romanian
Bhojpuri	Kannada	Russian
Bulgarian	Kazakh	Serbian
Burmese	Khmer	Sindhi
Cambodian	Korean	Sinhalese
Cantonese	Kurdish	Somali
Croatian	Lao	Spanish, Castilian
Czech	Latvian	Swahili

Danish	Lithuanian	Swedish
Dari	Luganda	Tagalog
Dhivehi	Malay	Taiwanese
Dutch	Malayalam	Tamil
English	Malaysian	Telugu
Estonian	Maltese	Thai
Farsi	Mandarin	Tigrinya
Finnish	Marathi	Turkish
French	Marwari	Ukrainian
Ganda	Mizo	Urdu
Georgian	Mongolian	Vietnamese
German	Nepali	Western Balkans
Greek	Norwegian	Zulu

Did Meta undertake any work with the intention of providing users cross-service communication between Meta Services and WhatsApp?

Meta responded 'no' to this question.

Was information shared by default between the following services when an account on one of these services was banned or suspended for CSEA?

Facebook information was shared with WhatsApp	Yes
Instagram information was shared with WhatsApp	Yes

Specify what information was shared by default when an account on one of these services was banned or suspended for CSEA

Meta stated that for severe violations of its child sexual exploitation, abuse and nudity policies, Meta shared specific indicators (which eSafety has chosen not to disclose) and ban reason with WhatsApp so as to enable a ban of the user's matched WhatsApp account. Meta stated that this may not have occurred in relation to users located in certain jurisdictions due to local privacy and other compliance obligations.

Parental controls

Provide the proportion of children's accounts on the following services that have parental supervision[132] enabled.

Service[133]	Proportion of 13 to 17 year olds accounts in Australia that had parental supervision enabled.	Proportion of 13 to 15 year olds accounts in Australia that had parental supervision enabled.
Facebook	Less than 1%	Less than 1%
Facebook Messenger	Less than 1%	Less than 1%
Instagram	1%	3.2%

[132] For example: https://about.fb.com/news/2023/06/parental-supervision-and-teen-time-management-on-metas-apps/

[133] Meta stated that the data provided in this table reflects the proportion of Australian children's accounts that had parental supervision enabled as at 15 December 2024. Meta noted that, for example, any users under the age of 16 who do not have parental supervision enabled will be subject to high safety and privacy settings as part of Teen Accounts (as they require parental permission to change to a less protective setting).

Microsoft

Key insights:

- Microsoft did not use any tools or alternative measures to detect CSEA livestreaming on Teams.

Proactive detection of CSEA

Was work undertaken to research, develop, source or implement tools to detect known CSEA images on OneDrive (material that is stored)?

Microsoft responded 'yes' to this question.

Specify the aims of the research and/or development conducted to detect known CSEA images on OneDrive (material that is stored)

Microsoft stated that OneDrive was undertaking development work to extend the use of hash-matching technologies to stored image and video content for Australian users.

Provide an explanation of alternative measures Microsoft Corporation took to detect known CSEA images on OneDrive (material that is stored)

Microsoft stated that it offered an in-product user reporting functionality for users to submit potential CSEA content for review.

Was work undertaken to research, develop, source or implement tools to detect known CSEA videos on OneDrive (material that is stored) and Outlook?

OneDrive (material that is stored)	Yes
Outlook	No

Provide details and specify the aims of the research and/or development conducted to detect known CSEA videos on OneDrive (material that is stored)

Microsoft stated that OneDrive was undertaking development work to extend the use of hash-matching technologies to stored image and video content for Australian users.

Provide an explanation of alternative measures Microsoft Corporation took to detect known CSEA videos on OneDrive (material that is stored) and Outlook

Microsoft stated that OneDrive did not implement hash-matching technologies when content was uploaded. When image or video content was shared, PhotoDNA and MD5 hash detection technologies were deployed to remove known content and suspend the user account.

Microsoft stated OneDrive offered an in-application user reporting functionality for users to submit potential CSEA content for review. If the reported content contained CSEA, the content was removed from the service and the user who shared the content was blocked from Microsoft services. The content owner was notified of the enforcement action, and the content was reported to NCMEC.

Outlook offered an in-application user reporting functionality for users to report email containing potential CSEA content in video format. If the reported content was confirmed to be CSEA content, the sending user's account was blocked if the sender was within in the Microsoft ecosystem. The sending user was notified of the enforcement action, and the content was reported to NCMEC.

Was any work undertaken to research, develop, source or implement tools to detect new CSEA material OneDrive (material that is stored), OneDrive (material that is shared), Outlook and Teams?

Microsoft stated that it is developing and piloting a program for Teams that was designed to identify and prevent potential new CSEA from being created or transmitted on live video calls made on the service. The end goal was for detection and mitigation of CSEA material in a timely manner, where Microsoft can detect, moderate, action, and report these accounts and individuals either before, during, or after the live video abuse occurs.

Provide an explanation of alternative measures Microsoft Corporation took to detect new CSEA material on OneDrive (material that is stored), OneDrive (material that is shared), Outlook and Teams

Microsoft stated that Teams used classifiers to detect 'adult' content. Although not designed or trained for CSEA, in all cases an image flagged by the classifier for nudity and routed for human review may have been determined to be new CSEA.

Microsoft stated that Teams, Outlook, OneDrive and Xbox utilised an in-product reporting flow where users could report CSEA content they encountered or a user who they believed was sharing CSEA. The potential CSEA content went through a human review and may have been determined to be new CSEA. If the reported content contained CSEA, the content was removed from the service and the user who shared the content was blocked from Microsoft services. In such cases, the content owner was notified of the enforcement action, the content was reported to NCMEC and a hash was created of the content and added to Microsoft's internal database of verified hashes.

Did Microsoft Corporation undertake any work to research, develop, source or implement tools to detect the CSEA livestreaming on Teams and Xbox?

Microsoft stated that it is developing and piloting a program for Teams that was designed to identify and prevent potential new CSEA from being created or transmitted on live video calls

made on the service. The end goal was for detection and mitigation of CSEA material in a timely manner, where Microsoft can detect, moderate, action, and report these accounts and individuals either before, during, or after the live video abuse occurs. Microsoft stated that Xbox did not provide video call or livestreaming functionality.

Provide an explanation of alternative measures taken to detect CSEA livestreaming on Teams and Xbox, and any other relevant context

Microsoft did not provide an explanation of alternative measures taken to detect CSEA livestreaming on Teams. Microsoft stated that there were significant jurisdictional and other conflicts associated with operating a global service for use by individuals in one country to communicate with individuals in other countries and that as such, Microsoft did not deploy classifiers or other automated content detection tools on video conferences held through Microsoft Teams. Microsoft stated that Xbox did not provide video call or livestreaming functionality.

Proactively detecting CSEA activity – grooming

Specify the languages the text and/or audio language analysis technology used to detect likely terms, abbreviations, codes and hashtags indicating grooming, operated in.

Community Sift:		
Arabic	Indonesian	Spanish
Chinese	Italian	Swedish
Danish	Japanese	Thai
Dutch	Korean	Turkish
Finnish	Polish	Ukrainian
French	Portuguese	Vietnamese
German	Romanian	
Hindi	Russian	
Internal proprietary tool A:		
English	Spanish	
Internal proprietary tool B:		
English	Spanish	

Did Microsoft Corporation undertake any work to research, develop, source or implement tools to detect grooming on Teams?

Microsoft responded ‘yes’ to this question.

Provide details, and specify the aims of the research and/or development conducted to detect grooming on Teams

Microsoft stated that Teams had invested time collaborating with several internal Microsoft teams to determine how Internal proprietary tool B could best be implemented on the service to detect grooming, with a focus on user messaging in Teams Communities.

Provide an explanation of alternative measures Microsoft Corporation took to detect grooming on Teams

Microsoft stated that Teams utilised an in-application product reporting functionality where users could report potential instances of grooming. The potential grooming content went through a human review. If a human moderator confirmed that Microsoft's grooming policy had been violated, the account of the underlying user was terminated and removed from the service, and the content owner was notified of the enforcement action.

Proactively detecting CSEA activity – sexual extortion

Specify the languages the text and/or audio language analysis technology used to detect likely terms, abbreviations, codes and hashtags indicating sexual extortion, operated in.

Community Sift:		
Arabic	Indonesian	Spanish
Chinese	Italian	Swedish
Danish	Japanese	Thai
Dutch	Korean	Turkish
Finnish	Polish	Ukrainian
French	Portuguese	Vietnamese
German	Romanian	
Hindi	Russian	
Internal proprietary tool A:		
English	Spanish	
Internal proprietary tool B:		
English	Spanish	

Did Microsoft Corporation undertake any work to research, develop, source or implement tools to detect sexual extortion on Teams?

Microsoft responded 'no' to this question.

Provide an explanation of alternative measures Microsoft Corporation took to detect sexual extortion of adults and children on Teams

Microsoft stated that Teams utilised an in-product reporting application 'Report a Concern' functionality where users could report potential instances of sexual extortion. The potential sexual extortion content went through a human review. If the reported content was determined to have violated Microsoft's sexual extortion policy, the content was removed from the service, the user who shared the content was blocked from Microsoft services, and the content owner was notified of the enforcement action.

Blocking URLs to known CSEA material

Was any work undertaken to research, develop, source or implement tools to block URLs linking to known CSEA on Outlook and Teams?

Microsoft responded 'no' to this question.

Provide an explanation of alternative measures Microsoft Corporation took to prevent URLs linking to known CSEA materials on Outlook and Teams

Microsoft stated that Outlook utilised an in-product reporting functionality 'Report a Concern' flow where users could report emails with URLs containing potential CSEA material. If the URL was determined to contain CSEA material, the account of the sender was terminated if the sender was in the Microsoft ecosystem.

Microsoft stated that Teams utilised an in-product reporting functionality where users could report URLs with potential CSEA material. If confirmed, the message containing the URL was taken down. Microsoft stated that Teams also planned to leverage a link reputation service that was being developed by a platform team within Microsoft.

Did Microsoft Corporation take and specific action based on the indicators to prevent recidivism?

Microsoft stated that new Microsoft account registrations were compared against accounts previously closed for sexual extortion of adults using indicators. These comparisons were used to identify similarities with the CSEA-associated accounts and limit re-registration by the same account holders.

Sharing information to reduce CSEA (and sexual extortion of adults)

Additional information

Microsoft stated that it had partnered with StopNCII.org and was piloting a program for Bing that involves accepting and scanning against hashes from StopNCII.org. Microsoft was actively

monitoring opportunities for improvement and ways to expand the program to other Microsoft services.

Resourcing to ensure safety

Languages that Microsoft's human moderators (employees) operated across		
OneDrive, Outlook and Teams:		
English	French	Portuguese
Spanish		
Xbox:		
English	Spanish	

Languages that Microsoft's human moderators (contractors) operated across		
OneDrive, Outlook and Teams:		
English		
Xbox:		
English	French	Portuguese
Spanish		

Additional information

Microsoft stated that it is developing and piloting a program for Teams that was designed to identify and prevent potential new CSEA from being created or transmitted on live video calls made on the service. The end goal is for detection and mitigation of CSEA material in a timely manner, where Microsoft can detect, moderate, action, and report these accounts and individuals either before, during, or after the live video abuse occurs.

Skype

Proactive detection of CSEA

Was any work undertaken to research, develop, source or implement tools to detect new CSEA material Skype?

Skype responded 'yes' to this question.

Provide details and specify the aims of the research and/or development conducted to detect new CSEA material on Skype

Skype stated that during the reporting period it has been developing and piloting a program designed to identify and prevent potential new CSEA from being created or transmitted on live video calls made on the service. The end goal of the program is for detection and mitigation of CSEA material in a timely manner, where Microsoft would be able to detect, moderate, action and report these accounts and individuals either before, during or after the live video abuse occurs.

Provide an explanation of alternative measures Skype took to detect new CSEA material

Skype stated that it utilised an in-product user reporting functionality where users could report CSEA content they encountered or a user who they believe was sharing CSEA. The potential CSEA content went through a human review to determine if it was new CSEA. If so, the content was removed from the service and the user who shared the content was blocked from Microsoft services. In such cases, the content owner was notified of the enforcement action, the content was reported to NCMEC, and a hash was created of the content and added to Microsoft's internal database of verified hashes.

Skype also stated that it used data analysis to prevent recidivism of users who have past CSEA related violations.

Did Skype undertake any work to research, develop, source or implement tools to detect the livestreaming of CSEA?

Skype responded 'yes' to this question.

Provide details, and specify the aims of the research and/or development conducted to detect the livestreaming of CSEA

Skype stated that during the reporting period, it has been developing and piloting a program designed to identify and prevent potential new CSEA from being created or transmitted on live video calls made on the service. The end goal of the program is for detection and mitigation of CSEA material in a timely manner, where Microsoft would be able to detect, moderate, action and report these accounts and individuals either before, during or after the live video abuse occurs.

Provide an explanation of alternative measures taken to detect CSEA livestreaming on Skype and any other relevant context

Skype referred to its response to alternative measures taken to detect known CSEA videos on Skype as well as the aims of the research and/or development conducted to detect the livestreaming of CSEA. Skype stated that there are significant jurisdictional and other conflicts associated with operating a global service for use by individuals in one country to communicate

with individuals in other countries and that, as such, Skype did not deploy classifiers or other automated content detection tools on video conferences held through Skype.

Proactively detecting CSEA activity – sexual extortion

Did Skype undertake any work to research, develop, source or implement tools to detect sexual extortion on the service?

Skype responded 'no' to this question.

Provide an explanation of alternative measures Skype took to detect sexual extortion of adults and children.

Skype stated that it utilised an in-product user reporting functionality where users could report content they encountered or report a user who they believed was attempting to engage in sexual extortion behaviour. The reported user who shared the content was blocked from Microsoft services. The content owner was then notified of the enforcement action.

Proactively detecting CSEA activity – grooming

Did Skype undertake any work to research, develop, source or implement tools to detect grooming?

Skype responded 'no' to this question.

Provide an explanation of alternative measures Skype took to detect grooming

Skype stated that it utilised an in-product user reporting functionality where users could report content they encountered or report a user who they believed was attempting to engage in grooming behaviour. The reported user who shared the content was blocked from Microsoft services. The content owner was then notified of the enforcement action.

Blocking URLs to known CSEA material

Was any work undertaken to research, develop, source or implement tools to block URLs linking to known CSEA on Skype?

Skype responded 'no' to this question.

Provide an explanation of alternative measures Skype took to prevent URLs linking to known CSEA materials

Skype stated that it provided in-product reporting functionality for users to report potential URLS linking to known CSEA material. The reported user was blocked from Microsoft services.

Information about the user and user's device were submitted to NCMEC and authorities when applicable.

Resourcing to ensure safety

Languages that Provider's human moderators (employees and contractors) operated across

Languages covered by Skype human moderators (employees):

English

Languages covered by Skype human moderators (contractors):

English

Snap

Key insights:

- Snap provided a warning message to Snapchat users who were listed as under 18 when they received contact from a stranger who had been reported by other users or who was from a region that person does not usually receive messages from, however 98.94% of users clicked 'okay' which led them to see the message. Only 0.61% of users reported or blocked the user instead of clicking through to read the message.
- Only 2.4% of Australian users whose accounts reflected that they were aged 13 to 17 were part of Snap's 'Family Centre' feature.

Proactive detection of CSEA

Did Snap Inc. undertake any work to research, develop, source or implement tools to detect known CSEA images on Snapchat?

Snap responded 'no' to this question.

Provide an explanation of alternative measures Snap took to detect known CSEA images on Snapchat, and any other relevant context

Snap referred to its responses in table 1 Snap stated that Snaps were novel photos or videos taken using Snapchat app's camera that, by definition, would be unlikely to result in any matches with hashes of known CSEA images.

Did Snap Inc. undertake any work to research, develop, source or implement tools to detect known CSEA videos on Snapchat?

Snap responded 'no' to this question.

Provide an explanation of alternative measures Snap took to detect known CSEA videos on Snapchat, and any other relevant context

Snap referred to its responses in table 2. Snap stated that Snaps were novel photos or videos taken using Snapchat app's camera that, by definition, would be unlikely to result in any matches with hashes of known CSEA images.

Did Snap Inc. undertake any work to research, develop, source or implement tools to detect the livestreaming of CSEA?

Snap responded 'no' to this question as it pertained to Video Chat on Snapchat.

Provide an explanation of alternative measures Snap took to detect livestreaming of CSEA on Snapchat, and any other relevant context

Snap stated that it did not have any alternative measures in place to detect CSEA. Snap stated that proactively monitoring private communications - including video calls – between individuals would be a breach of user trust and a violation of privacy.

Proactively detecting CSEA activity – grooming

Specify all the languages tools used to detect likely terms, abbreviations, codes, and hashtags indicating grooming on Snapchat

Snap stated that its text and/or audio language analysis tools operated in English, Arabic, Dutch, French, German, Norwegian, Spanish and Swedish during the reporting interval.

Proactively detecting CSEA activity – sexual extortion

Specify all the languages tools used to detect likely terms, abbreviations, codes, and hashtags indicating sexual extortion on Snapchat

Snap stated that its text and/or audio language analysis tools only operated in English, Dutch, French, German, Norwegian, Spanish and Swedish during the reporting interval. Preventing the spread and creation of CSEA - recommender systems.

What measures did Snap use to test and update its recommender system to contribute to the overall safety of Snapchat and avoid the risk of amplifying unlawful and harmful content?

Internal audits	Yes
External audits	Yes
Risk assessments	Yes
Impact assessments	Yes
a/b testing	Yes
Other	Pre-launch testing as well as a comprehensive annual review of recommender systems

How often were the tests to update recommender systems to contribute to the overall safety of Snapchat and avoid the risk of amplifying unlawful and harmful content performed?

Internal audit: Snap stated that its internal audit and risk team annually reviewed an inventory of significant algorithmic systems, including reviewing the frequency and nature of testing by the relevant engineering teams.

External audit: Snap stated that in accordance with Snap's obligations under Article 37 of the European Union Digital Services Act, Snap conducted an independent external audit annually to assess its compliance with the obligations in Chapter III of the EU DSA. These obligations included amongst others, the obligation to test and adapt Snap's content recommender system to ensure that unlawful and harmful content was not amplified.

Risk assessment: Snap stated that in accordance with Snap's obligations under Article 34(1) of the EU Digital Services Act, Snap conducted an annual risk assessment to identify, analyse and assess any systemic risks stemming from amongst others, the design or functioning of Snap's algorithmic systems. The risk assessment covered an in-depth assessment of the risks to and mitigations for child safety. It also included an in-depth assessment of the risks of dissemination of unlawful and harmful content on Snapchat and the mitigations Snapchat put in place to minimise such risks. Such content includes child sexual abuse material, hate speech, information related to the sale of prohibited products or services, terrorist content, content that infringes on intellectual property rights, adult sexual content, content regarding harassment & bullying, content that glorifies self-harm, including suicide, content relating to violent or dangerous behaviour, harmful false information, fraud and spam, and information related to other illegal activities.

Impact assessment: Snap stated that prior to deployment of any material change to Snap's recommender system, Snap's Product Legal team conducted a review as part of the Privacy and Safety by Design Review Process, and worked with related Engineering and Product teams to implement necessary changes.

A/b testing: Snap stated that it ran a/b testing for various aspects of Snapchat's content ranking model, some of which included safety improvements, about 2-5 times per month.

Pre-launch testing: Snap stated that it conducted model offline testing when appropriate. Offline testing aims to allow for quick, iterative model performance adjustments based on standard training sets. Snap conducted pre-launch offline testing using a population of Snap data.

Comprehensive review of recommender systems: Snap stated that it conducted a comprehensive annual review to centrally catalogue algorithmic systems that were significant to the functioning of Snapchat products as well as to safeguarding user safety and fundamental rights. This process was used to confirm understanding and documentation of significant algorithmic systems and review alignment of algorithmic systems with Snap's policies and obligations. Snap conducted this review annually.

Sharing information to reduce CSEA (and sexual extortion of adults)

Specify whether Snap took and used all available hashes from these databases or a subset. If a subset, specify what this consisted of, and explain why all hashes were not used.

Initiative	Did Snap Inc. take and use all available hashes, or a subset?	If a subset, specify what that subset consisted of	If a subset, explain why all hashes were not used
Take it Down	All	N/A	N/A
StopNCII	A subset	All except those marked as 'withdrawn' or 'deleted'.	Hashes marked as 'withdrawn' or 'deleted' by StopNCII were not considered reliable indicators in contrast with hashes that are labelled as 'active' by other initiative participants and thus more likely to be valid.

Preventing the spread and creation of CSEA – generative artificial intelligence

How did Snap Inc. ensure that Snapchat's generative AI features would not produce CSEA material?

Snap stated that its work to stop users from abusing Snapchat's generative AI features was aligned with legal requirements, which prohibited any person from intentionally attempting to generate child sexual abuse material, including through red-teaming.

Snap stated that it prohibited users from using its generative AI features to create content that violates its Terms of Service (including its Community Guidelines), which included a strict prohibition on CSEA material.

Snap stated that it used various mechanisms to prevent harmful and violative content (like CSEA material) from being generated with its products, including blocking words in user text prompts that may lead to the generation of harmful content, using image classifiers to detect harmful content and block generating an image based on that harmful content, and using prompt rewriting techniques to sanitise text prompt and help ensure generated images align with Snap's policies and guidelines.

Snap stated that it had conducted and continues to conduct AI Safety testing to prevent outputs that violate Snap's Community Guidelines, which includes CSEA material.

Snap stated that it had not attempted to produce images/symbols associated with CSEA using its generative AI features, as doing so could expose it to legal risks. However, Snap stated that it did conduct testing to prevent nudity and other forms of harm.

Explain why Snap did not take measures to ensure that known CSEA material was not included in the datasets used to train generative AI models available on Snapchat

Snap stated that it licensed generative AI models from OpenAI and Google to power several generative AI features available on Snapchat. OpenAI and Google have made public statements about their commitment to the Safety by Design Generative AI principles developed by Thorn and All Tech is Human. These principles aim to mitigate the risks generative AI poses to children and include a commitment to responsible source training datasets, and safeguard them from CSAM and CSEM.

Snap stated that Open AI stated that it detected and removed CSAM and CSEM from training data, and reported any confirmed CSAM to relevant authorities[134].

Snap also stated that Google made public statements that it was integrating both hash-matching and child safety classifiers to remove CSAM as well as other exploitative and illegal content from training datasets[135].

Explain why Snap did not take measures to ensure that new CSEA material was not included in the datasets used to train generative AI models available on Snapchat

Snap stated that it licensed generative AI models from Open AI and Google to power several generative AI features available on Snapchat. Snap stated that both Open AI and Google have made public statements about their commitment to detect and remove CSAM and CSEM from training data.

[134] Open AI, 2024, 'OpenAI's commitment to child safety: adopting safety by design principles', accessed 6 March 2025, URL: https://openai.com/index/child-safety-adopting-sbd-principles/
[135] Google, 'An update on our child safety efforts and commitments', 23 April 2024, accessed 6 March 2025, URL: https://blog.google/technology/safety-security/an-update-on-our-child-safety-efforts-and-commitments/

Resourcing to ensure safety

List all the languages that Snap Inc. human moderators (employees and contractors) operated across.[20]

Languages covered by Snap human moderator (employees):		
Arabic	Hebrew	Punjabi
Danish	Hindi	Romanian
Dutch	Italian	Russian
English	Japanese	Spanish
Finnish	Norwegian	Swedish
French	Polish	Turkish
German	Portuguese	

Languages covered by Snap human moderator (contractors):		
Arabic	Hebrew	Punjabi
Danish	Hindi	Romanian
Dutch	Italian	Russian
English	Japanese	Spanish
Finnish	Norwegian	Swedish
French	Polish	Ukrainian
German	Portuguese	Turkish

Expanded in-app warnings

What proportion of teenagers in Australia who received a warning message notifying them that the stranger has been reported by other users, or is from a region that the user does not normally receive messages, clicked on the following options:

Report or block (or equivalent)	0.61%[136]
'Okay' (or equivalent option that leads to them being able to message the user)	98.94%[137]

[136] Snap stated that this represents the percentage of times an Australian teenager under the age of 18 clicks 'Block or Report' when they see the warning message. Snap noted that: 0.45% of the time, the user will choose to ignore the warning message (that is, to not click 'Block or Report' or 'Okay') and leave the message on view. Snap stated that due to Snap's data retention policies, it only had data from 18 September to 15 December 2025.

[137] Snap stated that this represents the percentage of times an Australian teenager under the age of 18 clicks 'Okay' when they see the warning message. Snap noted that 0.45% of the time, the user will choose to ignore the warning message (that is, to not click 'Block or Report' or 'Okay') and leave the message on view. Due to Snap's data retention policies, Snap stated that it only had data from 18 September to 15 December 2025.

Are users who are reported frequently by other users, and send messages to users in regions that that user does not normally receive messages from, also flagged to Snap Inc. Trust and Safety staff for review, and potential bans?

Snap responded 'yes' to this question.

Parental controls

Provide the proportion of children's accounts in Australia that are part of Snap's 'family centre'[138] feature

Proportion of 13 to 17 year olds accounts in Australia that are part of Snap's 'family centre' feature	2.4%
Proportion of 13 to 16 year olds accounts in Australia that are part of Snap's 'family centre' feature	2.6%

[138] See: https://parents.snapchat.com/family-center

WhatsApp

Key insights:

- WhatsApp was heavily reliant on user reporting to detect CSEA (69% user reporting, 31% proactive detection), but WhatsApp had no specific category to report CSEA. When a report was made, WhatsApp only received the last 5 messages from the user/group reported to human moderators to review.
- WhatsApp did not share information of accounts banned with Facebook or Instagram at scale due to certain legal and technical restrictions on the sharing of WhatsApp user data with Meta.

Compared with previous response[139]

In our 2022 transparency report, eSafety reported that WhatsApp only reviewed 5 messages from a chat when a message was reported (compared to 30 in Facebook Messenger, also E2EE). WhatsApp did not make any improvement to its process and still only reviewed 5 messages when a message was reported. This means that WhatsApp's moderators are potentially missing important context when reviewing user reports, particularly in relation to grooming or sexual extortion.

Additionally, although Meta reported sharing Facebook information with WhatsApp in its response to eSafety's 2022 notice (and Meta reported sharing Instagram information with WhatsApp in its most recent periodic notice response), WhatsApp reported that information was not shared with Facebook or Instagram when an account was banned on WhatsApp for CSEA related reasons.

[139] A non-periodic notice was given to WhatsApp in 2022. Note that the questions asked and the reporting timeframes differed between notice rounds. These comparisons serve as a guide to how service responses may have changed since the last notice they received. Future comparisons will be observed through the following three periodic notice responses.

Proactive detection of CSEA

Did WhatsApp undertake any work to research, develop, source or implement tools to detect the livestreaming of CSEA?

WhatsApp responded 'no' to this question. WhatsApp stated that video calls on WhatsApp were end-to-end encrypted, which meant that WhatsApp could not access the contents of a video call. However, users were able to report other users on a video call.

Proactively detecting CSEA activity – grooming

Specify the languages the text and/or audio language analysis technology used to detect likely terms, abbreviations, codes and hashtags indicating grooming operated in.

List of languages used to train the internal proprietary tool A:

Afrikaans	Hausa	Polish
Albanian	Hebrew	Portuguese
Amharic	Hindi	Punjabi
Arabic	Hindi Romanised	Romanian
Armenian	Hungarian	Russian
Assamese	Icelandic	Sanskrit
Azerbaijani	Indonesian	Scottish Gaelic
Basque	Irish	Serbian
Belarusian	Italian	Sindhi
Bengali	Japanese	Sinhala
Bengali Romanised	Javanese	Slovak
Bosnian	Kannada	Slovenian
Breton	Kazakh	Somali
Bulgarian	Khmer	Spanish
Burmese	Korean	Sundanese
Catalan	Kurdish (Kurmanji)	Swahili
Chinese (Simplified)	Kyrgyz	Swedish
Chinese (Traditional)	Latin	Tamil
Croatian	Latvian	Tamil Romanised
Czech	Lithuanian	Telugu
Danish	Loa	Telugu Romanised
Dutch	Macedonian	Thai
English	Malagasy	Turkish
Esperanto	Malay	Ukrainian
Estonian	Malayalam	Urdu

Filipino	Marathi	Urdu Romanised
Finnish	Mongolian	Uyghur
French	Nepali	Uzbek
Galician	Norwegian	Vietnamese
Georgian	Oriya	Welsh
German	Oromo	Western Frisian
Greek	Pashto	Xhosa
Gujarati	Persian	Yiddish

Did WhatsApp undertake any work to research, develop, source or implement tools to detect grooming?

WhatsApp responded 'no' to this question.

Provide an explanation of alternative measures WhatsApp took to detect grooming.

WhatsApp stated that it did not use any language analysis technology specifically for the purpose of detecting grooming in private messaging, but did use other detection tools which could detect material and/or behaviour associated with grooming, for the purposes of enforcement.

Steps that WhatsApp took to minimise the risk of grooming in private messages include:

- It did not recommend or allow discovery of other users. A user needed to know a person's phone number to be able to contact them.
- The first time a user received a message from a non-contact, WhatsApp provided a 'first message experience', which included safety tools and tips, and clearly surfaced an option to block the non-contact.
- In addition, media received from non-contacts is blurred by default and not downloaded.
- WhatsApp deployed machine learning classifiers on reported messages to detect for violations such as CSAM solicitation and sexual extortion, which could be associated with grooming.

Proactively detecting CSEA activity – sexual extortion

Specify the languages the text and/or audio language analysis technology used to detect likely terms, abbreviations, codes and hashtags indicating sexual extortion operated in.

List of languages used to train the internal proprietary tool A:

Afrikaans	Hausa	Polish
Albanian	Hebrew	Portuguese
Amharic	Hindi	Punjabi
Arabic	Hindi Romanised	Romanian
Armenian	Hungarian	Russian
Assamese	Icelandic	Sanskrit
Azerbaijani	Indonesian	Scottish Gaelic
Basque	Irish	Serbian
Belarusian	Italian	Sindhi
Bengali	Japanese	Sinhala
Bengali Romanised	Javanese	Slovak
Bosnian	Kannada	Slovenian
Breton	Kazakh	Somali
Bulgarian	Khmer	Spanish
Burmese	Korean	Sundanese
Catalan	Kurdish (Kurmanji)	Swahili
Chinese (Simplified)	Kyrgyz	Swedish
Chinese (Traditional)	Latin	Tamil
Croatian	Latvian	Tamil Romanised
Czech	Lithuanian	Telugu
Danish	Loa	Telugu Romanised
Dutch	Macedonian	Thai
English	Malagasy	Turkish
Esperanto	Malay	Ukrainian
Estonian	Malayalam	Urdu
Filipino	Marathi	Urdu Romanised
Finnish	Mongolian	Uyghur
French	Nepali	Uzbek
Galician	Norwegian	Vietnamese
Georgian	Oriya	Welsh
German	Oromo	Western Frisian
Greek	Pashto	Xhosa
Gujarati	Persian	Yiddish

Blocking URLs to known CSEA material

Was any work undertaken to research, develop, source or implement tools to block URLs linking to known CSEA?

WhatsApp responded 'no' to this question.

Provide an explanation of alternative measures WhatsApp took to prevent URLs linking to known CSEA materials

WhatsApp stated that it was unable to access material on encrypted parts of the service, which meant that WhatsApp was unable to use technology to block URLs on these parts of the service.

However, WhatsApp stated that it did take measures to detect known child sexual exploitation and abuse material as set out in its previous responses on hash-matching. In addition, WhatsApp used classifiers to detect potential child sexual exploitation and abuse material in text (including URLs) on the following parts of the service (which are not encrypted):

- material reported by users
- title and description of Groups, Communities and Channels
- Channel posts.

Investigating user reports of CSEA

In order to identify CSEA, grooming and sexual extortion material contained within end-to-end encrypted messages on WhatsApp, how many messages between users could WhatsApp and/or its human moderators review in response to a user report?

WhatsApp stated that when a user reported another user or a group, WhatsApp received the last 5 messages from that user or group. Users could also report as many individual messages (or media) as appropriate. When an account or group was reviewed, all reported message content for that account or group (within 30 days) was available for review.

Did WhatsApp LLC undertake work to investigate increasing the number of messages sent to WhatsApp and/or its moderators to provide additional context to enable the identification of CSEA, grooming or sexual extortion material?

WhatsApp responded 'no' to this question.

Did WhatsApp LLC have a specific user reporting category for CSEA?

WhatsApp responded 'no' to this question.

If no, what steps did WhatsApp LLC take to ensure its user reporting options for CSEA were clear and readily identifiable, and enable WhatsApp to appropriately prioritise reports?

WhatsApp stated that a user was able to easily report another user or group by using WhatsApp's in-app reporting tool. When a user reported another user or group, they were not required to select a specific reporting category. This minimised the number of clicks required by the user and was intended to lower the barrier to reporting, especially for low-literacy users, and generally make reporting easier for the user. WhatsApp used technology to estimate the probability that a user report relates to a violation of WhatsApp's policies, which allowed WhatsApp to appropriately prioritise user reports for human review.

Did WhatsApp LLC obtain insights or data from Meta regarding the effectiveness of harm specific (including CSEA) user reporting options on Facebook and/or Instagram?

WhatsApp responded 'no' to this question.

Specify any other research or insights that WhatsApp LLC had regard to in determining not to implement a specific user reporting option for CSEA.

WhatsApp responded 'no' to this question.

Was information shared between the following services when an account was banned on WhatsApp for CSEA-related reasons?

WhatsApp information was shared with Facebook	No
WhatsApp information was shared with Instagram	No

If no, provide an explanation of alternative measures WhatsApp LLC took to share information with others with the purpose of preventing and dealing with CSEA activity and/or material and sexual extortion, and any other relevant context.

WhatsApp stated that it was unable to share information with Facebook and Instagram at scale due to certain legal and technical restrictions on the sharing of WhatsApp user data with Meta. WhatsApp stated that it shared information with Meta for the purposes of specific investigations, including those related to CSEA, and with a particular focus on networks of adversarial actors. In addition, WhatsApp shared such details with law enforcement if there was an imminent threat to life.

Did WhatsApp LLC undertake work with the intention of providing users cross-service communication with Meta services?

WhatsApp responded 'no' to this question.

WhatsApp's 'view once' feature

Were Android and iOS users prevented from being able to take screenshots of 'view once' content?

Android users	Yes
iOS users	Yes

WhatsApp included a link to more information on the 'view once' feature: https://faq.whatsapp.com/1077018839582332/

Resourcing to ensure safety

Languages covered by WhatsApp human moderators (employees):

WhatsApp did not provide a response.

Languages covered by WhatsApp human moderators (contractors):

Arabic	Hebrew	Spanish
Bahasa	Hindi	Urdu
English	Pashto	
Farsi	Portuguese	

eSafety.gov.au