# Basic Online Safety Expectations

## Summary of industry responses to the first mandatory transparency notices

**December 2022**

Australian Government | eSafety Commissioner

esafety.gov.au

# Contents

# 1. Executive summary

On 29 August 2022, the eSafety Commissioner (eSafety) gave non-periodic reporting notices (notices) to seven online service providers (providers) requiring each provider to report on its implementation of the Basic Online Safety Expectations (the Expectations) with respect to child sexual exploitation and abuse (CSEA). eSafety received responses from all the providers.

Notices were given under section 56(2) of the Online Safety Act 2021 (the Act). Under the Act, eSafety can publish summaries of the information received through the notices, making eSafety's powers relating to the Expectations a world-leading tool for lifting transparency and accountability from industry. The information obtained in response to these notices provides valuable insights that have not been forthcoming from voluntary initiatives to date, including in providers' own transparency reports.

The aim of this report is therefore to improve transparency and accountability of providers by providing better information about what they are actually doing – or not doing – to keep Australians safe, and to incentivise services to improve their safety standards.

The Commissioner noted when the notices were issued that:

> We have seen a surge in reports about this horrific material since the start of the pandemic, as technology was weaponised to abuse children. The harm experienced by survivors is perpetuated when platforms and services fail to detect and remove the content...
>
> We know there are proven tools available to stop this horrific material being identified and recirculated, but many tech companies publish insufficient information about where or how these tools operate, and too often claim that certain safety measures are not technically feasible.
>
> Industry must be upfront on the steps they are taking, so that we can get the full picture of online harms occurring and collectively focus on the real challenges before all of us. We all have a responsibility to keep children free from online exploitation and abuse.

Providers were asked specific questions about the tools, policies and processes they are using to address various forms of CSEA, such as the proliferation of online CSEA material, the online grooming of children, and the use of video calling and conferencing services to provide live feeds of child abuse.

This report shows significant variation in the steps being taken by providers to protect users and the wider Australian public. There is no common baseline, either between providers or even across a provider's own services. For example, while eSafety found some providers use well established 'digital fingerprinting' technology tools to identify images or videos previously identified as being CSEA material across all the services eSafety asked about, other providers use these tools on some of their services, but not others. These tools have an error rate of about 1 in 50 billion[1]. Until now, providers have not been open about these differences.

Some providers are checking for new or 'unseen' CSEA material, or using technology to detect potential grooming conversations, while eSafety was told by another provider that there is no technology good enough for either purpose. Most providers who were asked did not identify specific steps being taken to identify the abuse of children through live video calls, conferences or streams.

There is significant variation in the steps being taken to prevent recidivism (where users banned for previous abuse re-register with new accounts). Some providers report tracking extensive lists of indicators of recidivism, while others report only using a minimal number. There is also significant variation about what information is shared between a provider's own services to prevent banned users operating on multiple parts of a provider's products.

There are significant differences in the speed with which providers respond to user reports of child sexual exploitation, with responses varying from 4 minutes to 2 days (and 19 days where eSafety were told cases needed 're-review'). Some other providers have no reporting options at all within the app or service, requiring users to contact the provider via e-mail if they wish to complain about illegal or harmful activity on a service.

eSafety recognises that each provider is different, with different architectures, business models and user bases. This means an intervention, or use of specific tools on one platform, may not be proportionate on another.

However, where the notices have provided information of potential failure to implement the Expectations, eSafety will be engaging with those platforms to understand their plans and any obstacles to compliance.

Although the notices cover only a small number of providers, the information received represents a first step towards greater transparency. eSafety will issue further notices to additional providers on child sexual exploitation and abuse in 2023, before expanding to other harms to improve transparency over the coming months.

[1]Farid H (2019) '**Statement to the House Committee on Energy and Commerce Hearing: Fostering a Healthier Internet to Protect Consumers**', accessed 28 November 2022.

# 2. Glossary

- **The Act:** The Online Safety Act 2021.
- **Age assurance:** Measures to understand a user's likely age.
- **Age verification:** Measures to confirm a user's age.
- **Artificial intelligence (AI):** Generally considered to be digital technology which has the capability to exhibit human-like behaviour when faced with a specific task. AI can rapidly process data and spot patterns enabling these tools, in the context of this report, to support the moderation of online content.
- **Automated tools:** Digital technology used to sort data into categories automatically. In the context of this report, these tools are used to support content moderation actions and decisions.
- **Children:** Users below the age of 18 years.
- **Class 1 material:** Defined in section 106 of the Act by reference to the National Classification Scheme, includes material that is both sexually exploitative and that depicts or describes child sexual abuse.
- **CSEA:** child sexual exploitation and abuse. Unless specified, in an online context, this can refer to both material and activity (for example grooming).
- **CSEA material:** A broad category of material, normally referring to images and videos depicting the sexual abuse of a child, including sexual assault (Child sexual abuse material or 'CSAM'), as well as content that sexualises and is exploitative of a child, but that does not necessarily show the child's sexual abuse (Child sexual exploitation material or 'CSEM').
- **Dark web:** A layer of the internet composed of websites that are not indexed by search engines and can only be accessed using special networks. Often, the dark web is used by individuals who want to remain anonymous.
- **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 or young person for sexual activity at a later time.
- **Hash:** Dividing a digital file into small pieces and combining them to yield a concise numerical value that can be used to identify or match the original. May be referred to as a unique 'digital signature' or '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 hash of an image or video which is then compared against hashes of other photos to find copies of the same image or video.

- **Known CSEA material:** 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:** Use of artificial intelligence to assign a probability that text or conversations involve grooming.
- **Livestreamed CSEA:** The broadcasting of acts of sexual exploitation or abuse of children via webcam or video to people anywhere in the world, sometimes in exchange for payment.
- **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), any other form or any combination of forms. In the context of CSEA, 'material' typically refers primarily to images or video content.
- **NCMEC:** The US 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:** Material 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.
- **The Act:** The Online Safety Act 2021.
- **Recidivism:** In the context of this report, banned or suspended users re-registering to an online service with new details.
- **Video-calling 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.

# 3. The Basic Online Safety Expectations

The Basic Online Safety Expectations Determination 2022 sets out the Australian Government's expectations that social media, messaging, gaming, dating, file sharing services and other apps and websites will take reasonable steps to keep Australians safe. The current determination was registered on 23 January 2022.

Compliance with the Expectations is not mandatory, but eSafety may exercise powers under the Act to require providers to report on the steps they are taking to meet the Expectations. There are financial penalties for providers that do not respond to the questions asked in a notice.

eSafety can also issue a 'service provider notification' identifying whether a provider has implemented the Expectations. These are also known as 'statements of compliance' and 'statements of non-compliance'. eSafety's Regulatory Guidance, published in July 2022, noted that eSafety would not issue statements of non-compliance in 2022 'until further guidance on reasonable steps is published, other than in exceptional cases.' Therefore, this report aims to improve transparency and accountability, but does not reach a conclusion on whether individual providers are meeting the Expectations at this point.

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

# 4. Information about the first notices

## 4.1 Who received notices?

In August 2022, eSafety issued the first non-periodic reporting notices to seven providers under section 56(2) of the Act. The notices covered the period 24 January 2022 to 31 July 2022 and required each provider to report on its implementation of specific Expectations listed in the notices as they relate to CSEA.

While some providers have the same parent company, they are considered separate providers for the purposes of the Expectations so were given separate notices. The providers and services captured by the notices are set out below.

| Provider that received the section 56(2) notice | Services |
|---|---|
| **Apple Pty Ltd.** | iCloud email<br>iCloud<br>iMessage<br>FaceTime |
| **Meta Platforms, Inc.** | Facebook<br>Messenger<br>Instagram |
| **WhatsApp LLC** | WhatsApp |
| **Microsoft Corporation** | OneDrive<br>Outlook.com<br>Xbox Live<br>Teams |
| **Skype Communications S.à.r.l.** | Skype |
| **Omegle.com LLC** | Omegle |
| **Snap Inc.** | Snapchat |

The Expectations do not apply to some types of online services, including app stores, internet search services and device manufacturers, which is why those services were excluded from notices, despite some of the providers offering those services. Some subsidiary services were also excluded if eSafety decided they were not priorities, relative to the considerations in the Act.

In deciding which providers to issue a notice to, eSafety is required to consider several criteria specified in the Act, including:

- the number of complaints eSafety has received under the Act in relation to the service in the previous 12 months
- any previous contraventions of civil penalty provisions related to reporting on the Expectations
- any deficiencies in the provider's safety practices and/or terms of use
- whether the provider has agreed to give the Secretary[2] regular reports relating to safe use of their service
- any other matters the Commissioner considers relevant.

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

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

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

## 4.2 What questions did eSafety ask?

The notices required providers to respond to eSafety in the manner and form specified in the notices. This involved responding to a set of specific questions, using a template provided by eSafety. The questions were a mix of yes and no questions, and questions allowing free text answers or seeking specific data. eSafety's view is that targeted questions assist both the provider and eSafety. It ensures the provision of meaningful information and minimises the regulatory burden on respondents. eSafety also did not seek information that was readily accessible in the public domain.

Through answering the questions, providers were required to report on the specific steps they are taking to meet the relevant Expectations by detecting and preventing CSEA on their services.

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

[2]Secretary of the Department of Infrastructure, Transport, Regional Development, Communications and the Arts

An overview of the types of questions asked is contained in the following table, with the corresponding Expectations listed.

| Areas covered by Notices | Corresponding Expectations in the Determination |
|---|---|
| The extent to which to providers are deploying technology tools to identify child sexual exploitation and abuse content on different parts of their service, particularly such content that has already been confirmed and 'hashed' (digitally fingerprinted to allow for matches to be detected), as well as tools to detect 'new' or previously unseen content | Section 6 (Ensuring Safe Use)<br>Section 11 (Minimising Provision of Certain Material) |
| Steps taken to prevent and detect livestreamed child sexual exploitation and abuse | Section 6 (Ensuring Safe Use)<br>Section 11 (Minimising Provision of Certain Material) |
| Steps taken to prevent and detect the grooming of children | Section 6 (Ensuring Safe Use) |
| Steps to detect child sexual exploitation and abuse on end-to-end encrypted services | Section 6 (Ensuring Safe Use)<br>Section 8 (Encrypted Services)<br>Section 11 (Minimising Provision of Certain Material) |
| Steps taken to prevent services that allow users to remain anonymous being used for child sexual exploitation and abuse | Section 6 (Ensuring Safe Use)<br>Section 9 (Anonymous Services)<br>Section 11 (Minimising Provision of Certain Material) |
| Steps taken to prevent banned or suspended users from creating new accounts (recidivism) | Section 6 (Ensuring Safe Use)<br>Section 14 (Policies and Procedures to deal with complaints) |
| Availability of mechanisms for users to report child sexual exploitation and abuse content or activity | Section 6 (Ensuring Safe Use)<br>Section 11 (Minimising Provision of Certain Material)<br>Section 13 (Mechanisms to report and make complaints about certain material)<br>Section 14 (Policies and Procedures to deal with complaints)<br>Section 15 (Mechanisms to report and make complaints about breaches of terms of use) |

## 4.3 What responses were received?

All providers responded within the time required. Providers had 28 days to respond, or longer as agreed with eSafety. Two-week extensions were granted in some instances. The extensions recognised the fact that these were the first notices issued by eSafety.

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

This document provides a summary of the information that eSafety received from the seven providers. eSafety has published summary tables outlining information that providers reported in relation to the same issues, as well as individual summaries of providers' responses to their notices.

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

In particular, eSafety has determined that it is not in the public interest to publish specific indicators and signals that providers use to detect users seeking to commit crimes and cause harm, to prevent recidivism, or any details of new technologies where these are not in the public domain already. eSafety engaged with law enforcement agencies and other child safety experts to seek views on what kind of information would not be in the public interest to publish.

The following points should also be noted:

- The information provided in responses to the notices has not been verified by eSafety, although 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 point in time – the practices employed by the identified providers in the period 24 January 2022 to 31 July 2022 inclusive, or other periods within this timeframe as specified. Providers may have implemented changes to policies and processes since this information was provided.
- Where data has been provided it covers the period 24 January 2022 to 31 July 2022, unless otherwise stated.
- All data is global, unless otherwise stated.
- Terms used are defined in the glossary to this report, unless otherwise stated.

The information presented in this summary provides fresh insight into the efforts that these providers are making to address online CSEA. eSafety hopes that the information obtained from notices on the Expectations can be used by industry to discuss and address key challenges, and incentivise greater transparency in the future, including through their own voluntary disclosures. eSafety also hopes the information is used by researchers, academics, the media and public to scrutinise the efforts of industry, to improve accountability and encourage implementation of the Expectations.

## 4.5 What happens next?

- eSafety will use the information gathered from the first notices to build an understanding of industry practices, with a focus on improving transparency and accountability.
- eSafety will raise specific gaps and vulnerabilities with service providers that received notices to understand more about why certain aspects of the Expectations may not currently be complied with, and any future steps that are planned to ensure their services are implementing the Expectations.
- eSafety intends to issue further notices on CSEA in early 2023, before expanding to other acute harms, and welcomes input from all stakeholders on the areas where greater transparency is needed.
- eSafety's regulatory guidance, published in July 2022, sets out eSafety's planned approach to the exercise of its powers in respect of the Expectations. As well as further non-perioidic notices, in 2023 eSafety will also:
  - Issue the first periodic notices to begin tracking compliance with one or more metrics over time.
  - Publish any additional guidance necessary.
  - Begin issuing statements of compliance and/or non-compliance.
- eSafety will be making a decision shortly on whether to register the first set of industry codes, which were provided to eSafety by the industry associations on 18 November 2022. The Commissioner is required to consider the registration requirements under the Act, including whether these codes provide 'adequate community safeguards' in relation to matters of substantial relevance to the community. If they do not, eSafety can develop a standard. Information obtained through the first notices on the Expectations helps inform eSafety and, along with other information, provides helpful context regarding current industry practice and potential capabilities.

# 5.Transparency: Responses by issue

eSafety did not ask the same question of every provider. In some cases information was already published by a provider, or otherwise known to eSafety. In other examples, the question was not applicable because the relevant feature was not available on that service. Further, eSafety determined that some issues carried inherently higher risk than others and opted to prioritise questions about issues that it considered to have a higher risk of harm. This ensured that the regulatory burden was minimised, and eSafety's questions focussed transparency where it was needed the most.

Where providers reported on information that addressed the same or similar issues, eSafety has compiled that information in summary tables commencing at page 12. This enables greater transparency by setting out the actions that providers are taking side-by-side. It also provides a fuller understanding of how different providers within industry operate.

eSafety recognises that each provider is different – with different 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. Accordingly, when reviewing the tables, the nature of the service and the context in which the service operates should be taken into account. Information also reflects a point in time, and policies and processes may change.

eSafety has also explained in this section why it asked questions related to particular issues and provides a high-level overview of the technology available to industry.

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

## 5.1 Detecting previously confirmed ('known') images

Providers were asked about their use of tools to detect images containing known CSEA content. 'Known' images are those that have been previously assessed and confirmed as CSEA. For example, the US based National Center for Missing and Exploited Children houses a non-governmental organisation (NGO) database of child sexual abuse material triple vetted by expert analysts from different global children's organisations.

There are a variety of tools available to identify matches of these known images. PhotoDNA is one of the most widely used tools. It was developed in 2009 by Microsoft and Dartmouth College and has since been made freely available to qualifying organisations. It is an example of a 'hash matching' tool, which creates a unique digital signature (known as a 'hash') of an image which is then compared against signatures (hashes) of other photos to find copies of the same image. PhotoDNA's error rate is reported to be 1 in 50 billion.[3] These tools play a vital role preventing the ongoing re-victimisation of children and adults, whose images otherwise circulate endlessly online. These tools protect both the privacy of the general population given their high accuracy rates, and that of victims.

[3]Farid H (2019) '**Statement to the House Committee on Energy and Commerce Hearing: Fostering a Healthier Internet to Protect Consumers**', accessed 28 November 2022.

eSafety asked about the detection of known CSEA material as an example of a reasonable step providers could take to meet the expectations contained in:

- Section 6(2) of the Determination: that providers 'proactively minimise the extent to which material or activity on the service is unlawful or harmful'.
- Section 11 to minimise the extent to which the 'class 1' material, which includes CSEA, is provided on the service.

The following information was given by service providers regarding the use of hash matching tools to identify images containing known CSEA material in response to the notices:

| Provider | Services | Uses image hash matching tools on these services Yes / No | Names of tools used |
|---|---|---|---|
| **Microsoft** | OneDrive (stored content that is shared)<br>Outlook.com<br>Xbox Live<br>Teams (when not E2EE) | Yes | • PhotoDNA[4]<br>• MD5 |
| | OneDrive (stored content that is not shared) | No | N/A |
| **Skype** | Skype<br>• Messaging (when not E2EE[5]) | Yes | • PhotoDNA<br>• MD5 |
| **Snap** | Snapchat:<br>• Discover<br>• Spotlight<br>• Direct Chat | Yes<br>(Snap noted that it scans images uploaded from device to Snapchat on Snap map, Chat, Discover and Spotlight) | • PhotoDNA |
| | Snapchat<br>• Snaps[6] | No<br>(Snap noted that Snaps are taken 'live' so content is unlikely to involve known CSEA imagery) | N/A |
| **Apple** | iCloud email | Yes | • PhotoDNA |
| | iMessage (E2EE by default[7])<br>iCloud | No | N/A |
| **Meta** | Messenger (when not E2EE)<br>Facebook<br>Instagram<br>Instagram direct (when not E2EE) | Yes | • PhotoDNA<br>• PDQ[8]<br>• Proprietary matching technologies<br>• SSN++ (for Facebook and Instagram only) |
| **WhatsApp** | WhatsApp (which is E2EE by default[9]) | Yes, for<br>• user profile and group images<br>• images found in user reports | • PhotoDNA<br>• Other proprietary technologies. |

[4]Microsoft, 'PhotoDNA', 2022, accessed 28 November 2022 [5]End-to-end encryption on Skype is an opt-in feature. [6]Snaps are end-to-end encrypted by default [7]iMessage is end-to-end encrypted by default [8]Meta, 'Open-Sourcing Photo- and Video-Matching Technology to Make the Internet Safer', 2019, accessed 28 November 2022 [9]WhatsApp is end-to-end encrypted by default

## 5.2 Detecting previously confirmed ('known') videos

Providers were asked about whether they were deploying hash matching tools to detect videos containing known CSEA content. PhotoDNA for Video, Google's CSAI Match and Facebook's TMK+PDQF are examples of existing tools, made available to organisations. As video content is increasingly ubiquitous online, these tools have become more important than ever to assist in preventing the dissemination of CSEA material. The Tech Coalition, a CSEA industry organisation, has announced work to improve the standardisation of video hashing tools used across industry.

As with known images, eSafety asked about the detection of known CSEA videos in relation to section 6(2) of the Determination, and section 11.

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

| Provider | Services | Uses video hash matching tools on these services Yes / No | Names of tools used |
|---|---|---|---|
| **Microsoft** | Xbox Live<br>Teams (when not E2EE[10]) | Yes | • PhotoDNA for Video[11] |
| | OneDrive (both when content shared and not shared)<br>Outlook.com | No | N/A |
| **Skype** | Skype Messaging (when not E2EE) | Yes | • PhotoDNA for Video |
| **Snap** | Snapchat:<br>• Discover<br>• Spotlight<br>• Direct Chat | Yes<br>(Snap noted that they scan video uploaded from device to Snap on Snap map, Chat, Discover and Spotlight) | • CSAI Match[12] |
| | Snapchat:<br>• Snaps | No<br>(Snap noted that Snaps are taken 'live' so content is unlikely to involve known imagery) | N/A |
| **Apple** | iMessage (which is E2EE by default)<br>iCloud<br>iCloud email | No | N/A |
| **Meta** | Messenger (when not E2EE)<br>Facebook<br>Instagram<br>Instagram direct (when not E2EE) | Yes | • eSafety did not ask Meta to name the tools used |
| **WhatsApp** | WhatsApp (which is E2EE by default) | Yes, for<br>• Videos found in user reports | • An internally developed tool for hashing and matching video frames |

[10]End-to-end encryption is only available on Microsoft Teams for enterprise [11]Microsoft, '**How PhotoDNA for Video is being used to fight online child exploitation**', 2018, accessed 28 November 2022 [12]Google, '**Child safety toolkit**', accessed 28 November 2022

## 5.3 Detecting new material

Hash matching tools can only 'match' against previously identified and confirmed CSEA material and seek to prevent its ongoing dissemination. But steps also can be taken to prevent the sharing of CSEA material when it is first created, and before it has been identified and included in a database. This may occur through the use of artificial intelligence ('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 material, as well as material that does not contain CSEA content, in order to identify the markers of content depicting abuse. An example of this technology is Google's Content Safety API[13] or Thorn's classifier, which Thorn reports has a 99% precision rate.[14]

eSafety asked about the detection of new content in relation to section 6(2) of the Determination, and section 11.

In response to the notices, the following information was given by providers regarding tools being used to detect images containing new CSEA material:

| Provider | Services | Uses tools on these services to identify new CSEA images Yes / No | Names of tools used |
|---|---|---|---|
| **Microsoft** | Xbox Live<br>Teams (when not E2EE)<br>One Drive (content shared and not shared)<br>Outlook.com | No | While no CSEA tools are used to detect new images, Microsoft say that OneDrive, Teams and Xbox Live use classifiers to detect nudity, which also picks up some CSEA content |
| **Skype** | Skype Messaging (when not E2EE) | No | N/A |
| **Snap** | Snapchat:<br>• Discover<br>• Spotlight<br>• Direct Chat<br>• Snaps | No | N/A |
| **Meta** | Facebook<br>Instagram<br>Messenger (when not E2EE)<br>Instagram direct (when not E2EE) | Yes | Google's Content Safety API<br><br>Internally built CSAM classifier & child sexualization classifier (Facebook and Instagram only) |
| **WhatsApp** | WhatsApp (which is E2EE by default) | Yes, for<br>• User profile, Group images;<br>• Images found in user reports | Google's Content Safety API<br><br>Internal classification model |

[13]Google, '**Child safety toolkit**', accessed 28 November 2022 [14]Thorn (2020), '**How Safer's detection technology stops the spread of CSAM**', accessed 28 November 2022.

## 5.4 Detecting CSEA in livestreams or video calls

The Australian Institute for Criminology has published analysis of offender chat logs showing the use of 'popular platforms to arrange and watch the sexual abuse of children live'.[15] During the United Kingdom's Independent Inquiry into Child Sexual Abuse hearings various industry members were questioned about and acknowledged the risks posed by livestreamed CSEA on their services.[16] The Australian Centre to Counter Child Exploitation has highlighted that Australian children as young as 8 years old are being coerced into performing livestreamed sexual acts by online predators, who often record and share the videos elsewhere – including on the dark web – and sexually extort victims into producing even more graphic content.[17] Sometimes this is in exchange for payment. The International Justice Mission has described livestreamed CSEA material as 'live crime scenes happening on tech platforms.'[18]

Detecting CSEA content in a live video is more technically challenging than detecting still images, given the volume of content transmitted. However, there are examples of steps being taken to deploy technology to detect harmful content in live videos.

For example, the provider Yubo has published information[19] about how they detect harmful content in live video, noting:

> We monitor all livestreams, through first-of-its-kind video and audio moderation to detect inappropriate behaviour, and intervene in real-time when we see our Community Guidelines being broken. This might include users putting hate speech in a Live title, using drugs or weapons, posing in their underwear or discussing self-harm behaviours during the livestream.

The safety tech company SafetoNet has also published[20] information about their 'SafeToWatch' tool, providing 'a real-time video threat detection tool... to automatically detect and block the filming and viewing of CSAM [Child Sexual Abuse Material].' Beyond detection technologies providers could also put in place measures such as safety prompts or age assurance measures.

All providers that eSafety asked about steps to detect livestreaming drew a distinction between 'livestreaming' and 'video calls' or 'video conferences'. However, eSafety's understanding from their responses to the notice questions is that the providers are neither taking action to detect CSEA in livestreams (insofar as any of these could be regarded as livestreaming services) or taking action to detect CSEA in video calls or conferences.

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

In response to the notices, the following information was given by providers regarding detecting CSEA material in livestreaming, or video calls/video conferences:

[15]Australian Institute for Criminology (2021), "Live streaming of child sexual abuse: An analysis of offender chat logs", accessed 12 December 2022. [16] IICSA (2019) 'IICSA Inquiry - Internet Hearing', accessed 12 December 2022 [17]Australian Federal Police (2021) 'AFP warn about fast growing online child abuse trend', accessed 29 November 2022. [18]International Justice Mission (2022) 'IJM Submission to Public Consultation on the Draft Consolidated Industry Codes of Practice for the Online Industry (Class 1A and 1B Material)', accessed 29 November 2022. [19]Yubo, (n.d.) 'Safety Hub', accessed 29 November 2022. [20]Camera Forensics (2022) 'Understanding real-time threat detection and the SafeToWatch mission', accessed 29 November 2022.

| Provider | Services | Measures in place to detect CSEA in livestreams or video calls/conferences – Yes/No | Tools used |
|---|---|---|---|
| **Microsoft** | Teams | No | N/A |
| **Skype** | Skype video calls | No | N/A |
| **Apple** | FaceTime[21] | No | N/A |
| **Omegle** | Omegle | Yes | Hive AI[22] |

## 5.5 Detecting grooming

Grooming is the process of building a relationship with a child in order to sexually abuse or exploit them. This increasingly involves children or young people being tricked or coerced into sexual activity on webcams or into sending sexual images through online services. Detecting grooming on online services presents technical challenges, particularly as every service will have a different user base, linguistic style, type of conversation and language. Context is also important, for example language indicative of grooming may have an innocent explanation – such as poor humour – which context can reveal.

Despite these challenges, there are a growing number of examples of technologies to support providers in identifying grooming, while protecting privacy. Microsoft supported Project Artemis in 2020, which involved a collaboration with The Meet Group, Roblox, Kik and Thorn, to develop a grooming detection tool.[23] The child safety organisation Thorn has also since made an anti-grooming tool available to providers. A number of providers also have internal language tools to detect likely grooming. Some tools provide users with warning messages and prompts if a communication looks suspicious. Others support content moderation efforts by prioritising accounts for further human review.

A recent anonymised survey by the WeProtect Global Alliance and the Tech Coalition found that 37% of tech companies surveyed currently use tools to detect the online grooming of children.[24]

eSafety asked about the detection of grooming in relation to section 6(2) of the Determination.

[21]FaceTime is end-to-end encrypted by default [22]Hive, '**Cloud-based AI solutions for understanding content**', 2022, accessed 29 November 2022 [23]Microsoft (2020) '**Microsoft shares new technique to address online grooming of children for sexual purposes**', accessed 29 November 2022. [24]WeProtect (2021) '**Findings from WeProtect Global Alliance/Tech Coalition survey of technology companies**', accessed 29 November 2021.

In response to the notices, the following information was given by providers regarding the language analysis tools used to detect grooming activities:

| Provider | Services | Uses language analysis tools on these services Yes / No | Names of tools used |
|---|---|---|---|
| **Microsoft** | Xbox Live | Yes | 'Multiple tools and processes' using text analysis to develop a confidence score |
| | Outlook.com<br>Teams (when not E2EE)<br>OneDrive (for both content shared and not shared) | No | N/A |
| **Skype** | Skype Messaging (when E2EE enabled and not enabled) | No | N/A |
| **Snap** | Snapchat:<br>• Direct chat<br>• Snaps | No | N/A |
| **Apple** | iMessage (E2EE by default) | No | N/A |
| **Meta** | Instagram direct (when not E2EE)<br>Messenger (when not E2EE) | Yes<br>(For prioritising reported messages) | An internally built machine learning classifier that combines internal language analysis technology with other information |
| **Omegle** | Moderated Text Chat | Yes | Omegle's automated scanning system for text chats, which monitors for certain patterns including grooming |

## 5.6 Detecting recidivism

eSafety investigators regularly see the same offenders create multiple new accounts, even after they have been banned by a platform. In one recent instance, a child who had been sexually exploited online reported 22 accounts to eSafety, all of which had been used to harass and threaten the child. Each account displayed common characteristics, indicating the accounts were likely to have been created by one offender, or multiple offenders who were operating together, to perpetrate the online harm. eSafety also sees the same offenders operating on different services provided by the same parent company, despite being banned on one service.

eSafety has chosen not to publish the specific signals or indicators that providers use to prevent banned users re-registering to use services, to avoid them from being used to avoid detection. Instead eSafety has tried to capture the scale, or range, of indicators used. This is an imprecise metric, as some indicators will be more important than others and some providers use those indicators more proactively and rigorously than others. However, eSafety's view is that, in general, providers that are looking for a wider range of indicators will have a better chance of preventing the re-registration of banned users. eSafety has provided a number of providers with intelligence reports and guidance about how their platforms are being weaponised and the steps they can take to tackle these issues and remediate future harm.

eSafety asked about recidivism in the context of section 6 (2), section 11, as well as:

- **Section 14 (2)**'s expectation that providers will 'take reasonable steps to ensure that penalties for breaches of its terms of use are enforced against all accounts held or created by the end-user who breached the terms of use of the service'.

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

| Provider | Services | Steps to prevent recidivism | Number of indicators |
|---|---|---|---|
| **Microsoft** | Xbox Live<br>Outlook.com | Yes | Several |
| **Skype** | Skype Messaging (when E2EE enabled and not enabled) | Yes | Several |
| **Snap** | Snapchat | Yes | Several |
| **Meta** | Facebook<br>Instagram | Yes | Multiple |
| **WhatsApp** | WhatsApp (which is E2EE by default) | Yes | Minimal |
| **Omegle** | Omegle | Yes | Minimal |

For providers with multiple subsidiary services, there are steps that can also be taken to ensure that users banned on one subsidiary service for CSEA related activities are not able to use a provider's other services. The steps that providers can take may depend on the legal structure of subsidiaries in some instances.

| Provider | Services | Cross-service ban – Yes/No |
|---|---|---|
| **Microsoft** | Xbox Live<br>Teams (when not E2EE)<br>One Drive (content shared and not shared)<br>Outlook.com | Yes<br>(The Microsoft account is closed preventing access to all other Microsoft services accessed with a Microsoft account) |
| **Skype** | Skype | Yes<br>(The Microsoft account is closed preventing access to all other Microsoft services accessed with a Microsoft account) |
| **Meta** | Facebook<br>Instagram<br>Messenger | In specific circumstances |
| **WhatsApp** | WhatsApp | No<br>(In relation to sharing information with Meta's other services) |

## 5.7 Responding to user reporting

When illegal content such as CSEA material is reported by a user it should be actioned quickly to prevent ongoing or new harm. User reporting is a key safety measure, as reflected in the following expectations:

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

In response to the notices, the following information was given by providers regarding the median time for user-reported CSEA material to be actioned (e.g. content removed, user banned, or other content moderation decision taken).

| Provider | Services | Median time for user reported CSEA material to be actioned |
|---|---|---|
| **Microsoft** | Xbox Live | 1 day |
| | OneDrive (links shared) | 2 days |
| | Teams | 2 days[25] |
| **Skype** | Skype | 2 days[26] |
| **Snap** | Snapchat | 4 minutes |
| **Meta** | Facebook<br>Messenger | 75 minutes |
| | Instagram<br>Instagram Direct | 41 minutes |
| **WhatsApp** | WhatsApp | 26-29 hours |

In 2020 the Canadian Centre for Child Protection published a review[27] of different providers' user reporting options, highlighting the importance of easy to access in-service options, with specific options to flag that a report is for CSEA material, so it can be prioritised appropriately.

Most providers have relatively easy-to-find information about how users can report illegal and harmful content, and some have specific options for CSEA, which eSafety regards as best practice. However, eSafety asked Apple and Omegle to confirm their user reporting options, as eSafety could not find this information available prominently on the relevant services.

| Provider | Services | In service user reporting option | Additional comments by provider |
|---|---|---|---|
| **Apple** | iCloud email<br>iCloud<br>FaceTime<br>iMessage | No | Apple commented that users can report via the abuse@apple.com email address and via Apple support. |
| **Omegle** | Omegle | No | Omegle commented that users can report via the safety@omegle.com email address. |

[25]Microsoft and Skype's original response to the relevant questions in their notices indicated that the median time to respond was 19 days for Skype and Teams. Microsoft subsequently informed eSafety that 'this only accounted for certain queues used by our content moderators, where re-review is required; not all queues of user reported concerns were accounted for.' Microsoft and Skype's amended responses to the questions in the notices stated that the median time to respond from 24 January 2022 through 31 July 2022 was actually 2 days. [26]Ibid [27]Canadian Centre for Child Protection (2020) 'Reviewing child sexual abuse material reporting functions on popular platforms', accessed 29 November 2022.

# 6. Transparency: Individual provider responses

# 6.1 Apple summary

## Overview

Apple Pty Ltd was asked about its iCloud email, iCloud, iMessage, and FaceTime services.

### Questions about detecting known and new child sexual exploitation and abuse material

In relation to questions about hash matching for known child sexual exploitation and abuse (CSEA) material, Apple confirmed it was using PhotoDNA on the following services to detect known images:

| Hash matching tools for known CSEA images are used | Hash matching tools for known CSEA images are not used |
|---|---|
| iCloud email | iCloud |
| | iMessage |

Apple provided the following information regarding the use of hash matching for known CSEA videos:

| Hash matching tools for known CSEA videos are used | Hash matching tools for known CSEA videos are not used |
|---|---|
| None of the services eSafety asked about | iMessage |
| | iCloud |
| | iCloud email |

Apple was also asked about alternative measures taken to proactively minimise known CSEA material, but did not provide any additional information.

### Question about whether Apple intends to scan for CSEA content uploaded to iCloud Photos, as it previously announced and then delayed

Apple stated, 'While we do not comment on future plans, Apple continues to invest in technologies that protect children from CSEA'.

In answer to a follow up question from eSafety seeking a full response, Apple reiterated the same answer.

The background which informed this question was that in August 2021, Apple announced plans 'to detect collections of illegal, known CSAM [child sexual abuse material] images stored on Apple servers in iCloud Photos libraries, while not learning any information about non CSAM images.'[28] Apple described the feature as 'a privacy-preserving, hybrid on-device/server pipeline to detect collections of CSAM images being uploaded to iCloud Photos.'[29]

[28]Apple (2021) Security threat model review of Apple's child safety features, accessed 29 November 2022. [29]Apple (2021) Security threat model review of Apple's child safety features, accessed 29 November 2022.

At the time Apple stated that 'the possibility of any given account being flagged incorrectly is lower than one in one trillion'.[30]

Apple subsequently delayed the feature, and stated 'we have decided to take additional time over the coming months to collect input and make improvements before releasing these critically important child safety features.'[31]

On 7 December 2022, Apple announced that:

'We have further decided to not move forward with our previously proposed CSAM detection tool for iCloud Photos.'[32]

Apple also announced on 7 December 2022 that it was end-to-end encrypting iCloud backups.

### Question about ability of users to report CSEA activity within iMessage, FaceTime, iCloud and iCloud email

Apple confirmed that there are no user reporting options within iMessage, FaceTime, iCloud or iCloud email.

Apple commented that users can make complaints of CSEA material via Apple support or via email to abuse@apple.com.

### Question about the techniques or language analysis used to detect potential online grooming on iMessage

Apple confirmed that it does not use any language analysis technology to detect grooming on iMessage.

Apple highlighted its recent 'Communications Safety' in Messages intervention, which 'introduced tools to warn children when receiving or sending photos that contain nudity.'

In answer to a follow-up question, Apple clarified that this intervention was aimed at a number of online safety risks including grooming, and helped to combat grooming in 'two phases of the grooming process':

1. When the groomer shares images that contain nudity, Apple's 'Communication Safety' tool identifies the content and warns the child of the risks of viewing this material; and
2. If a child does decide to share an image that contains nudity, Apple's 'Communication Safety' tool identifies the content and warns about the risks of sending the image.

[30]Apple (2021) Security threat model review of Apple's child safety features, accessed 29 November 2022. [31]The Guardian (2021) Apple delays plans to scan cloud uploads for child sexual abuse images, accessed 29 November 2022. [32]Wired (2022), 'Apple Kills Its Plan to Scan Your Photos for CSAM. Here's What's Next', accessed 12 December 2022.

Apple stated that in both cases, the 'Communication Safety' tool provides resources to the child that helps them 'evaluate the situation, receive guidance on how to stay safe online, enable them to talk with an adult from either their direct circle or a hotline, and ultimately report the situation to the appropriate authority.' Apple added:

> Given that the objective of a groomer is to befriend, normalize sharing imagery containing nudity and/or sex between adults and children, and then coerce a child in sharing nude images, Communication Safety seeks to combat this outcome. Additionally, Apple provides informative notifications to users on the danger of grooming and the importance of notifying proper legal authorities at its occurrence.

Apple referred to additional tools such as the ability to block users and measures for child accounts, as well as 'the abuse@iCloud.com reporting mechanism'.

### Question about measures in place to prevent the livestreaming of CSEA on FaceTime

Apple confirmed that it does not have any measures in place to detect livestreaming of CSEA on FaceTime.

Apple made a distinction between 'livestreaming' and FaceTime 'video calls'. FaceTime was described as a 'private way to make video calls', which is 'limited by design to closed groups with no more than 33 participants'.

Apple stated that 'FaceTime is not designed for large live-streaming experiences,' and 'As such, Apple does not have any measures to detect the live-streaming of content on FaceTime.'

In response to a clarifying question from eSafety about steps to detect CSEA on either livestreaming or video calls, Apple reported that it 'instead, focuses on other measures such as, for example the ability to report CSAM materials via Apple Care and providing informational notifications to users on the steps to provide law enforcement with instances of CSAM on Apple services.'

### Question about the proportion of family accounts with children that have enabled safety notifications within iMessage for children's accounts, either globally or in Australia

Apple reported that it 'does not collect sufficient data to determine what proportion of family accounts with children have enabled this feature.'

### Questions about risk assessments when making policy or design changes to iMessage, FaceTime, iCloud or iCloud email

#### Conducting risk assessments

Apple confirmed that it conducts safety risk assessments when it makes policy or design changes to these services. Apple added that it 'considered online safety risks where relevant when making material policy, feature, or design changes' and that 'online safety risks represented one set of risks taken into account in Apple's decision making processes'.

### Steps to mitigate risks identified and ensure safety by design

Apple reported that during the period covered by the notice, 'online safety risks were considered' and that 'there were also a range of other risks that were considered and balanced on relevant decisions including privacy and security risks'.

Apple added that it 'sought to achieve a balanced and reasonable position when taking all relevant considerations into account.'

When asked how safety risks are evaluated alongside other types of risk, such as those identified in a privacy impact assessment, Apple responded that it is currently reviewing and supplementing its safety by design and risk assessment processes in light of Australia's Online Safety Act. Apple added, 'further changes to such processes will be informed by the industry codes of practice... given these are likely to include more specific requirements regarding risk assessments and safety by design.'

## Question about the use of metrics to internally assess efficacy of interventions to detect CSEA

Apple stated that it does not have internal metrics in place.

Apple added that it 'sees less value in collecting metrics that will not provide meaningful insights into the efficacy of measures – for example, Apple does not believe that complaints metrics provide a good basis for understanding the efficacy of measures.'

Apple further noted that:

> Complaints metrics can be influenced by a range of factors – including changes in reporting mechanisms and publicity around particular issues – and cannot be tracked in a manner that is tied to the efficacy of interventions.

# 6.2 Meta Summary

## Overview

Meta Platforms, Inc. was asked about its Facebook, Messenger, Instagram and Instagram Direct services.

### Questions about known and new child sexual exploitation and abuse material

In relation to questions about tools used to detect known CSEA material, Meta confirmed that it uses hash matching tools to identify known images, if end-to-end encryption is not enabled:

| Hash matching tools for known CSEA images are used | Names of the tools |
|---|---|
| Messenger (if E2EE not enabled) | PhotoDNA, PDQ, proprietary matching technologies |
| Facebook | SSN++, PhotoDNA, PDQ, proprietary matching technologies |
| Instagram | SSN++, PhotoDNA, PDQ, proprietary matching technologies |
| Instagram Direct (if E2EE not enabled) | PhotoDNA, PDQ, proprietary matching technologies |

Meta previously announced that it was open-sourcing a video matching / hashing tool, TMK+PDQF.[33] Meta confirmed to eSafety that it is using tools to detect known video CSEA material on Messenger, Facebook, Instagram, and Instagram Direct, if end-to-end encryption is not enabled.

Meta also confirmed it was using tools to detect new CSEA material if end-to-end encryption is not enabled, and provided the following details. Meta noted that there was some variability globally in the use of these tools to detect CSEA originating from certain jurisdictions due to 'local privacy and other compliance obligations'.

| Tools are used to detect new CSEA | Names of the tools |
|---|---|
| Messenger (if E2EE not enabled) | Google's Content Safety API |
| Facebook | Internally built CSAM classifier and child sexualization classifier; Google's Content Safety API |
| Instagram | Internally built CSAM classifier and child sexualization classifier; Google's Content Safety API |
| Instagram Direct (if E2EE not enabled) | Google's Content Safety API |

[33]Meta (2019), Open Source Photo Video Matching, accessed 12th December 2022.

## Questions about the reasonable steps taken to proactively minimise known and new CSEA material

Meta identified further steps it takes to proactively minimise known and new CSEA, outlining five main categories.

**Actively discouraging behaviour that may lead to sharing of CSEA material**

Meta reported that the following steps were in place to discourage sharing of CSEA:

- When a user is identified as having shared 'viral, meme child exploitative content' on one of Meta's services, Meta sends a safety alert to the user. The safety alert informs the user about the harm that can be caused and warns the user that the material is against Meta's policies and that there are legal consequences for sharing the material. When CSEA material is removed, Meta noted that it reports the material to the US based National Center for Missing and Exploited Children (NCMEC) and bans the account that promotes this material.* The material is also stored as a hash to prevent re-upload.
- The number of messages that can be forwarded on Messenger is limited to five people or group chats at a time.
- A user that searches for terms connected with child sexual exploitation is shown a pop-up notice that reminds the user that child sexual exploitation is illegal and directs them to 'offender diversion resources.'
- Removing Facebook profiles, Pages, groups and Instagram accounts that are dedicated to sharing otherwise innocent images of children alongside captions, hashtags or comments that indicate the sexualisation of children in the images. Meta noted that content that isn't explicit and doesn't depict child nudity is harder to define. However accompanying text can provide context.

**Additional proactive detection efforts**

- Meta identified that it has expanded efforts to detect and remove networks that violate its child exploitation policies and noted that the work is similar to its 'efforts against coordinated inauthentic behaviour and dangerous organizations.'
- Meta also noted that on its Facebook and Instagram services, it uses 'a rule-based classifier to detect known violating child sexualisation images' in order to remove them.

**Preventing inappropriate contact between adults and teens**

- When a 'teen' (which Meta clarified referred to a person under the age of 16 years) creates an account with Facebook or Instagram, the account defaults to a private or 'friends only' account.
- A 'pop-up safety notice' appears to a teen when messaging with an adult they may not know on Instagram Direct and Messenger. Meta identifies that the notice educates the teen to be cautious and empowers them to take action before responding to a message.

*Meta does not by default ban accounts across all its services, as noted in answers to questions below.

- Meta said it uses machine learning to analyse behavioural data 'across our platforms' to identify inappropriate interactions between an adult and teen. Meta stated that these processes do not include the scanning of private messages but do include the ability to identify indicators of potentially harmful intent by an adult toward a teen.
- On Facebook and Instagram, it has processes to try to prevent adults that have been identified as potentially harmful from discovering, searching for and connecting with teens and interacting with teens' content.
- On Instagram, Meta noted that an adult is prevented from sending a message to a person under 18 that does not follow them.

In a separate section, eSafety asked Meta if an adult can message a child on Messenger who they are not connected with ('friends' with). eSafety asked for a "yes or no" response. Meta responded 'yes'.

Meta provided additional context, adding that 'at a high level, adults cannot send messages on Messenger to teens outside of their broader social network.' Whether or not a teen falls within the adult's broader social network depends on various factors which Meta outlined in its report. Meta also identified that 'Teens who are messaging with a potentially suspicious adult also receive a pop-up notice encouraging them to block or report if something doesn't seem right.'

**Providing tools for users to report CSEA material**

- Meta said that reports involving children are prioritised.
- Meta reported that it consulted with child safety experts to make it easier to report content that violates child exploitation policies by making reporting tools easier to find, expanding the surfaces that users can specifically report content involving a child and reducing the steps involved.
- Additionally, Meta identified that it points teens towards reporting at key moments, such as when blocking another account.

**Contributing to research that aims to understand and prevent sharing of CSEA material**

- Meta reported that it has worked with experts on child exploitation, including NCMEC, to develop a research-backed taxonomy that may assist in identifying when a person has malicious intent in sharing CSEA material. Meta reported that its work in this area was ongoing.

## Question about identifying additional CSEA material

Meta was asked whether it reviews all user content when an account is confirmed to have shared CSEA material using technical tools or human moderators.

Meta reported that its automated image and video matching systems seek to review all available images and videos. Meta subsequently added that this only applies where content is not end-to-end encrypted. When known CSEA is detected on an account, those systems continue to search for other known CSEA on the same account on Facebook, Messenger, Instagram and Instagram Direct.

Meta said a human review may also be applied in circumstances where additional indicators of harm are identified about an account or a report about an account is received relating to potential abuse.

### Questions about the techniques or language analysis used to detect potential online grooming on Messenger and Instagram Direct

Meta confirmed that it uses language analysis tools on Messenger and Instagram Direct, but stated that these were only used to prioritise user reports.

Meta added that it uses 'An internally built machine learning classifier that combines internal language analysis technology with other information, such as information about the sender of the message(s), to rank user reports based on the likelihood that the reported message content is a grooming conversation.'

### Question about user reporting on Messenger's 'Secret Conversations'

Meta was asked why 30 recent messages on Messenger's end-to-end encrypted 'Secret Conversations' service are sent for review when a user reports a conversation. eSafety noted that in comparison Meta's WhatsApp service only sends five messages when an end-to-end encrypted WhatsApp message is reported by a user.[34]

Meta said that the choice to use 30 messages was because Messenger 'has currently sought additional message content specifically to enforce against harms that require further context from the conversation.'

Meta added that 'decisions on the amount of user-generated content to be reported from encrypted [end-to-end encrypted] conversations will vary by service.' Meta said that this can be informed by a range of factors 'including user expectations, the harms being enforced against and the specifics of the enforcement system.'

### Questions about the steps and indicators used to tackle recidivism on Facebook and Instagram

Meta identified interventions that it has put in place to prevent recidivism on Facebook and Instagram, which rely on multiple different indicators. eSafety has chosen not to publish the specific number or type of indicators used.

Meta said that if a user is banned on one service for child sexual exploitation and abuse related breaches, information is shared between Facebook (including Messenger) and Instagram and vice versa in certain circumstances. eSafety has chosen not to publish additional information about these circumstances to avoid this being exploited.

### Question about sharing information with WhatsApp about users banned on Facebook or Instagram

Meta said Facebook user information is shared with WhatsApp for the purpose of preventing unlawful or harmful material and activity, but noted 'this may not occur in relation to users located in certain jurisdictions due to local privacy and other compliance obligations.'

In answer to a follow-up question from eSafety, Meta reported that Instagram does not share information with WhatsApp for these purposes.

[34]WhatsApp (n.d.), About blocking and reporting contacts, accessed 29 November 2022.

## Question about user reporting

Meta was asked about the average time to respond to user reports of CSEA content. Meta provided the following information.

| Median time to action user reports of CSEA (e.g. content removed and reported, user banned, or other content moderation decision taken) | |
|---|---|
| Facebook & Messenger | 1.25 hours* |
| Instagram & Instagram direct | 0.69 hours* |

*Meta said that data covers the period 24 Jan 2022 to 31 July 2022. User reports use labels other than CSEA, but which are closely related.

## Question about the time CSEA has been present on the service when it is detected

| Median time CSEA has been present on the service when it is detected (either by technology or user reports) and removed | |
|---|---|
| Facebook & Messenger | 5.83 hours* |
| Instagram & Instagram direct | 0.04 hours* |

*Meta said that data covers the period 24 Jan 2022 to 31 July 2022. User reports use labels other than CSEA, but which are closely related.

## Questions about risk assessments when making policy or design changes

**Conducting risk assessments**

Meta reported that it conducts safety risk assessments when it makes policy or design changes to its services. It also confirmed that it has processes in place to review and assess risks, including safety risks, with respect to proposed policy and/or design changes on Facebook, Messenger and/or Instagram.

In addition to having an internal policy team dedicated to safety issues, with cross-functional teams' input, Meta reported that it also consults with external experts.

**Steps to mitigate risks identified and ensure safety by design**

Meta reported that it incorporates the following principles:

- If a potential risk is identified, cross-functional teams work on 'mitigations to implement to address the identified potential risks.' Meta noted that mitigation steps taken will vary in each case, depending on the nature of the risks that have been identified and an assessment of how best to address those risks.
- After the proposed policy or design change(s) are evaluated and mitigations are identified, agreed upon and implemented, Meta reported that it may track metrics or other relevant indicators to help evaluate how the service is being used with the mitigations in place.

When asked how safety risks are evaluated alongside other types of risk, such as those identified in a privacy impact assessment, Meta responded that safety risks are evaluated in addition to privacy impact assessment processes and 'typically in close proximity to any such privacy impact assessments.'

Meta also noted that safety risks are sometimes assessed during privacy impact assessments. Meta said that in some circumstances privacy controls or other privacy-centric mitigations can be implemented in an attempt to address safety risks.

### Question about the use of metrics to internally assess efficacy of interventions to detect CSEA

Meta confirmed that metrics are in place on Facebook, Instagram and Messenger and pointed to their Community Standards Enforcement Report.[35]

[35]Meta, (2022), 'Community Standards Enforcement Report: Q3 2022 Report', accessed 29 November 2022.

# 6.3 WhatsApp Summary

## Overview

WhatsApp LLC was only asked about the WhatsApp service.

### Questions about known and new child sexual exploitation and abuse material

WhatsApp reported that it uses PhotoDNA and other proprietary technologies to identify known CSEA images.

WhatsApp noted that it uses an internally developed tool for hashing and matching video frames for detecting known CSEA material in video.

WhatsApp reported that it also uses Google's Content Safety API and an internal classification model to identify new CSEA material.

It said these tools were used on the following content and surfaces which are not end-to-end encrypted:

- user profile and group images
- images and videos found in user reports
- group subject and description.

**Known and new CSEA**

WhatsApp reported that when it becomes aware of 'child exploitation imagery' (CEI) on the platform, it bans the accounts involved, removes the imagery, and reports them, along with associated account details, to the National Center for Missing and Exploited Children (NCMEC), as required by US law.

WhatsApp noted that it is prepared to respond to valid law enforcement requests, and that it had received feedback from law enforcement that its efforts had assisted in rescuing victims of child abuse.

### Questions about the reasonable steps taken to proactively minimise new CSEA material

WhatsApp provided further information, including:

- WhatsApp noted that users cannot search for unconnected people or groups, as a phone number is needed in order to connect. WhatsApp reported that the first time a user gets a message from outside their address book, WhatsApp provides a notification asking if the user wants to block or report the sender.
- WhatsApp commented that when an unknown individual sends a user an image, this is blurred by default, allowing for blocking and/or reporting the sender without seeing the image.
- WhatsApp reported that it gives users control to decide who can add them to groups and WhatsApp limits the number of chats a user can forward a message to at once, to help limit the spread of harmful viral content.

### Question about the indicators used to evaluate whether a group is sharing CSEA

eSafety referred to information published by WhatsApp on its website[36] about the evaluation of 'group information and behavior for suspected CEI sharing', and asked for the indicators used.

WhatsApp highlighted that it does not have access to the contents of the message due to end-to-end encryption, unless that message is reported by a user.

WhatsApp confirmed that it does evaluate group information and behaviour for suspected child exploitation imagery. WhatsApp also confirmed that it evaluates unencrypted surfaces in groups, including group images and group subject and description.

WhatsApp also stated that it uses other classifiers and human reviewers to detect a range of behavioural signals in groups and provided additional detail on this to eSafety, which eSafety has decided should not be published.

### Questions about the steps and indicators used to tackle recidivism

WhatsApp identified interventions it has put in place to prevent recidivism which rely on a minimal number of different indicators. eSafety has decided not to publish the specific number or type of indicators used.

WhatsApp reported that if a user is banned from WhatsApp for a violation related to CSEA, this information is not shared with Facebook or Instagram (services owned by Meta, the same parent company as WhatsApp) for the purpose of preventing unlawful or harmful material and activity.

### Questions about WhatsApp's 'view once' feature

WhatsApp was asked about its 'view once' feature, which is described on the WhatsApp website as a feature that allows users to 'send photos and videos that disappear from your WhatsApp chat after the recipient has opened them once.'[37]

In response to a question about how often CSEA was identified in user reports of 'view once' content, WhatsApp identified that during the period 1 July 2022 to 31 July 2022, CSAM material had been identified 1,211 times globally from user reports of 'view once' content.

In response to a question about whether WhatsApp conducted a safety risk assessment to understand whether the 'view once' feature was likely to increase the risk of child sexual exploitation and abuse, WhatsApp responded that it had 'conducted an integrity review of the 'view once' feature prior to its launch which resulted in specific changes to facilitate reporting and enforcement'. It also identified that following the launch of this feature 'WhatsApp has monitored reporting and ban rates so as to ensure the continuing efficacy of our integrity systems.'

[36]WhatsApp, (2021), '**How WhatsApp helps fight child exploitation**', accessed 29 November 2022. [37]WhatsApp, (n.d.), '**About view once**', accessed 29 November 2022.

WhatsApp was also asked to outline its decision-making process undertaken to determine not to implement a feature that notifies users when a screenshot is taken of 'view once' media, as other services such as Facebook and Messenger have. In response, WhatsApp reported that it is planning to enable screenshot blocking, which its user experience testing of WhatsApp users indicated was 'more helpful' than notifications and is testing this feature, and intends to 'roll it out to users soon.'

### Question about user reporting

WhatsApp was asked about the average time to respond to user reports of CSEA content. WhatsApp provided the following information.

| Median time to action user reports of CSEA (content removed and reported, user banned, or other content moderation decision taken) | |
|---|---|
| WhatsApp | 26 – 29 hours* |

*WhatsApp reported that data covers the period between 1 June 2022 to 31 August 2022. WhatsApp noted that due to data retention policies it only stores the data for a rolling 90-day period.

### Question about detection of CSEA

WhatsApp was asked what proportion of the global 300,000 accounts[38] it bans monthly for child sexual exploitation are proactively detected by WhatsApp as opposed to those reported by users. WhatsApp provided the following information, which it reported was for the period 1 July 2022 to 31 July 2022 and therefore represented a 'snapshot in time and it may change'.

| Detection Method | Proportion of banned accounts |
|---|---|
| Proactive detection<br>(WhatsApp did not receive a report against the user indicating child sexual exploitation that directly led to the user being banned by WhatsApp) | 73% |
| Reactive detection<br>(WhatsApp received a report against the user indicating child sexual exploitation that directly led to the user being banned by WhatsApp) | 27% |

### Question about checking message content for malware

WhatsApp has previously identified publicly that it checks messages for 'suspicious links' that may lead to 'malicious sites', and that these checks take place on users' devices.[39] In response to a question about whether WhatsApp performs any form of checks on any other content for suspicious content or malware, for example in a gif, file, or photo, WhatsApp answered 'no'.

[38]WhatsApp (2021) 'How WhatsApp helps fight child exploitation', accessed 29 November 2022. [39]WhatsApp (n.d.) 'About suspicious links', accessed 29 November 2022.

WhatsApp added that:

> WhatsApp does not have access to the contents of the message (unless that message is reported to us by a user). For this reason, we are unable to perform any form of check on any other content (for example in a gif, file, or photo) for suspicious content or malware, unless it is provided to us via a user report.

WhatsApp added that 'Suspicious link detection on WhatsApp is a simple rule-based function that occurs on the message recipient's app.'

WhatsApp further stated:

> This feature is hardcoded into the WhatsApp app, which neither suppresses the link nor reports it out to a third party. This process happens in the context of an end-to-end encrypted chat, where WhatsApp has no ability to access the message content. Further, unlike the detection of CSAM from unencrypted and reported content, WhatsApp does not detect malicious links by matching them against a given database of known harmful content.

### Questions about risk assessments when making policy or design changes

WhatsApp reported that it conducts safety risk assessments when making policy or design changes.

WhatsApp explained that it considers safety and integrity risks at multiple stages in the product development process including when planning to update features, building the updates and launching those updates. It also noted that retrospective assessments are conducted following the launch.

In addition, WhatsApp identified that their Integrity team, comprised of engineering, product management, data science, design, research and other functions, also 'focuses on building safety and integrity features within the product.'

WhatsApp reported that 'the Integrity team evaluates proposed change(s), assesses potential risks posed by the proposed change(s), and identifies appropriate mitigations to implement to address any identified risks' and provides support and advice to other teams on building updates with 'safety at the forefront.'

**Steps to mitigate risks identified and ensure safety by design**

WhatsApp reported that it incorporates 'key aspects of Safety by Design principles in its integrity approach' which it identified as including the following:

- If a potential risk is identified, cross-functional teams work on 'mitigations to implement to address the identified potential risks.' WhatsApp noted that mitigation steps taken will vary in each case, depending on the nature of the risks that have been identified and an assessment of how best to address those risks.
- After the proposed policy or design change(s) are evaluated and mitigations are identified, agreed upon and implemented, WhatsApp reported that it may track metrics or other relevant indicators to help evaluate how the service is being used with the mitigations in place.

When asked how safety risks are evaluated alongside other types of risk, such as those identified in a privacy impact assessment, WhatsApp responded that safety risks are evaluated in addition to privacy impact assessment processes and 'typically in close proximity to any such privacy impact assessments.'

WhatsApp also noted that safety risks are sometimes assessed during privacy impact assessments. WhatsApp said that in some circumstances privacy controls or other privacy-centric mitigations can be implemented in an attempt to address safety risks.

### Question about the use of metrics to internally assess efficacy of interventions to detect CSEA

WhatsApp confirmed that it had metrics in place. WhatsApp reported that this involved tracking the number of accounts banned for child sexual exploitation and abuse.

# 6.4 Microsoft Summary

## Overview

Microsoft Corporation was asked about its OneDrive, Teams, Outlook, and Xbox Live services.

### Questions about known and new child sexual exploitation and abuse material

In relation to questions about hash matching for known child sexual exploitation and abuse (CSEA) material, Microsoft stated that PhotoDNA and MD5 are used on images sent on the following services:

| Hash matching tools for known CSEA images are used | Hash matching tools for known CSEA images not used |
|---|---|
| Links to content in OneDrive (shared with others) | Content stored in OneDrive (not shared) |
| Consumer version of Teams (which is not end-to-end encrypted) | |
| Outlook.com | |
| Xbox Live | |

Microsoft stated that PhotoDNA for Video is used on videos sent on the following services:

| Hash matching tools for known CSEA videos are used | Hash matching tools for known CSEA videos are not used |
|---|---|
| Xbox Live | Outlook.com |
| Consumer version of Teams (which is not end-to-end encrypted) | OneDrive – either stored content (not shared), or links (shared) |

Microsoft reported that it does not have specific automated tools to detect new CSEA material:

| Tools are used to detect new CSEA imagery | Tools are not used to detect new CSEA imagery |
|---|---|
| None of the services eSafety asked about | Outlook.com |
| | OneDrive – either stored content (not shared), or links (shared) |
| | Xbox Live |
| | Teams |
| | Skype |

Regarding the detection of new CSEA content, Microsoft noted that classifiers are used to detect 'adult' content on OneDrive, Teams and Xbox Live. Microsoft report that in some cases, an image flagged for nudity and sent to a human moderator may be identified to be new CSEA material.

## Questions about steps to identify whether an account that has previously shared CSEA material contains other CSEA material

| Does Microsoft review all of the content shared by an account when CSEA is detected? | |
|---|---|
| Outlook.com | No |
| OneDrive (links shared) | No |
| Xbox Live | Yes<br>Microsoft also noted that 'When Microsoft becomes aware of CSEA material on the Xbox Live service, it conducts human review of all user content in the Xbox environment associated with that account.' |
| Teams | No |

Microsoft also noted in another section of their response that users' access to Xbox Live, Teams, OneDrive and Outlook.com is connected and controlled by the user's Microsoft account. When an account is closed for a CSEA violation, this therefore prevents them from accessing any of these services.

## Questions about the techniques or language analysis used to detect potential online grooming on Xbox Live

In answer to a question relating to the use of language analysis tools used to detect online grooming on Xbox Live, Microsoft highlighted current and future activity.

**Current activity to detect potential online grooming**
Microsoft reported that it uses 'multiple tools and processes' to detect grooming. The techniques use 'text analysis to develop a confidence score'. Microsoft added that 'when a confidence score threshold (set by the service) is reached, a human investigator reviews the data to make a final determination of whether the Xbox grooming policy was violated.'

If violations are confirmed, Microsoft reports the activity to the US based National Center for Missing and Exploited Children (NCMEC) and blocks the violating account from all Microsoft services. Microsoft provided further details which eSafety has decided not to publish.

**Future activity to detect potential online grooming**
Microsoft reported that it has work underway to expand current grooming detection processes on Xbox Live to include sexual extortion, a 'distinct violation which may sometimes occur in tandem with grooming activity.' Microsoft highlighted in further correspondence that these measures should not be considered to mean that no measures are currently taken to address sexual extortion on Xbox or other named services.

Microsoft also reported that a recent collaboration was initiated by its Xbox team, within the Tech Coalition industry group, which will explore the development of an industry standard for grooming detection. Microsoft highlighted that, 'this is illustrative of Microsoft's history of cooperation with other service providers to promote safety on their services.'

### Question about measures in place to prevent the livestreaming of child sexual exploitation and abuse on Microsoft Teams

Microsoft reported that measures were not in place to detect CSEA in live video conferences (or video calls) on Teams.

Microsoft commented that:

> As 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, Microsoft does not deploy classifiers or other automated content detection tools on video conferences held through Microsoft Teams.

Microsoft made a distinction between 'livestreaming', which eSafety had asked about, and 'real-time video conferences'. It identified that livestreams are 'open more broadly (to large numbers, or the general public) and involve a different level of interactivity', and real-time video conferences which it referred to as 'interpersonal communications'.

Microsoft noted that some jurisdictions consider that:

> conference bridges (including video conferences) to be interpersonal communications. Similar to a telephone call, they often cannot by law applicable to multi-jurisdiction communications be intercepted by others without express permission of the users. Other applicable laws may prohibit other forms of access to such communications.

### Questions about steps to prevent recidivism for CSEA related breaches on Xbox Live and Outlook.com

Microsoft identified interventions that it has put in place to prevent recidivism on Xbox Live and Outlook.com, which rely on several different indicators. eSafety has chosen not to publish the specific number or type of indicators used.

Microsoft noted that if a user is banned on one service for CSEA-related breaches, the user will be barred from accessing all Microsoft services, including Skype, a separate service also provided by Microsoft.

## Question about user reporting

Microsoft was asked about the average time to respond to user reports of CSEA content. Microsoft provided the following information.

| Median time to action user reports of CSEA (content removed and reported, user banned, or other content moderation decision taken) | |
|---|---|
| Teams | 2 days* |
| OneDrive | 2 days* |
| Outlook.com | Microsoft commented that no reporting history was availabe |
| Xbox Live | 1 day |

*Microsoft noted that the figures for Teams represent a combined average for Skype and Teams Consumer as reported data does not delineate between these services.

Microsoft's original response to the relevant question in their notice indicated that the median time to respond was 19 days for Teams. Microsoft subsequently informed eSafety that 'this only accounted for certain queues used by our content moderators, where re-review is required; not all queues of user reported concerns were accounted for.' Microsoft's amended response to the question in the notice stated that the median time to respond from 24 January 2022 through 31 July 2022 was actually 2 days.

Microsoft provided additional information, stating that for the period July-December 2021, over 99% of content actioned for CSEA on Microsoft's consumer services, including Skype, was detected through automated detection (scanning) and not through user reports.

Microsoft also added that additional resources in its newly established Digital Trust and Safety engineering team are dedicated to improving automated detection of CSEA material.

## Questions about risk assessments when making policy or design changes to its services

**Conducting risk assessments**

Microsoft reported that it conducts safety risk assessments when it makes policy or design changes to its services and that all changes must comply with and be assessed against the Microsoft Digital Safety Standard, which was launched in May 2022.

Microsoft also reported that it conducted a pilot of the Digital Trust and Safety Partnership's (DTSP) 'Safe Framework' in early 2022 and is working to roll the Framework out to all Microsoft services once the Safe Framework has been finalised.

**Steps to mitigate risks identified and ensure safety by design**

When asked about the steps it takes to mitigate risks identified as part of its risk assessment process, Microsoft responded with the following:

> Services are reviewed against the Microsoft Digital Safety Standard and the results are discussed with applicable senior leadership (CVP level) to align on remediation and resourcing.

Microsoft also reported that it used the DTSP Safe Assessment questionnaire, 'as a framework for examining the people, processes, and technology that contribute to managing content – and conduct-related risks', and pointed to the DTSP Safe Assessments Report for more information.

When asked how safety risks are evaluated alongside other types of risk, such as those identified in a privacy impact assessment, Microsoft responded that:

> Pursuant to the Microsoft Privacy Standard, privacy reviews are conducted to identify, assess, and mitigate potential privacy risks from the collection, storing and sharing of personal data when system capabilities or processes are being designed. Content moderation and other safety interventions often involve data processing and, as such, would be included in a data protection impact assessment.

Microsoft also highlighted that other 'service-specific' risks may require consultation with subject-matter experts in areas such as cybersecurity, lawful access, telecommunications, Responsible AI, accessibility, and corporate standards amongst others.

### Question about the use of metrics to internally assess efficacy of interventions to detect CSEA

Microsoft confirmed that metrics are in place on OneDrive, Outlook, Xbox Live and Teams and commented that a key Digital Safety Standard program objective was to use data to demonstrate how well content moderation systems are functioning, to ensure continuous improvement. Microsoft added that:

> Microsoft's Digital Safety Office, engineering and Responsible AI teams are working together to develop assessment methods for evaluating the measures currently in place to determine whether they are operating as intended and mitigating the intended risk(s).

# 6.5 Skype summary

## Overview

Skype Communications S.à r.l. was only asked about its Skype service.

### Questions about known and new child sexual exploitation and abuse material

In relation to questions about hash matching for known child sexual exploitation and abuse (CSEA) material, Skype confirmed that it uses the hash matching tools PhotoDNA and MD5 for images, and PhotoDNA for Video on video, on Skype Messaging when it is not end-to-end encrypted.

Skype stated that it does not use specific automated tools to detect new CSEA material on Skype.

### Question about steps to identify whether an account that has shared CSEA material contains other CSEA material

Skype stated that it does not review all user content when CSEA is detected.

### Question about the techniques or language analysis used to detect potential online grooming on Skype

Skype stated it does not use any language analysis tools to detect grooming.

### Questions about measures in place to prevent the livestreaming of child sexual exploitation and abuse (CSEA) on Skype and how many user reports of livestreamed CSEA have been received

Skype reported measures were not in place to detect CSEA on video calls made through Skype.

Skype commented that:

> As 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, Microsoft does not deploy classifiers or other automated content detection tools on video calls made through Skype.

Skype made a distinction between 'livestreaming', which eSafety asked about, and 'real-time video calls'. It identified that livestreaming is 'open more broadly (to large numbers, or the general public)', and real-time video calls are 'interpersonal communications'.

Skype added that:

> Similar to a telephone call, they [interpersonal communications] often cannot by law applicable to such multi-jurisdictional communications be intercepted by others without express permission of the users. Other applicable laws may prohibit other forms of access to such communications.

In response to a question about how many user reports of livestreamed CSEA Skype has received, Skype responded it 'did not identify any reports of CSEA within video conferences during the period 24 January 2022 to 31 July 2022', based on a manual review. Skype added that Microsoft does not have a separate data collection for this information.

### Question about steps to prevent recidivism on Skype

Skype identified interventions that it has put in place to prevent recidivism, which rely on several different indicators. eSafety has chosen not to publish the specific number or type of indicators used.

In another part of their response Skype reported that 'Microsoft use data analysis to prevent users who have been banned for past CSEA-related violations from creating new accounts with different details.'

Skype also noted in another section of their response that users' access to Skype is connected and controlled by the user's Microsoft account. When an account is closed for a CSEA violation, this therefore prevents them from accessing any service accessed with a Microsoft account.

### Question about user reporting

Skype was asked about the average time to respond to user reports of CSEA content. Skype provided the following information.

| Median time to action user reports of CSEA (content removed and reported, user banned, or other content moderation decision taken) | |
|---|---|
| Skype | 2 days* |

*Skype noted that the figures for Skype represent a combined average for Skype and Teams Consumer (a separate service provided by the same parent company) as reported data does not delineate between these services.

Skype's original response to the relevant question in the notice indicated that the median time to respond was 19 days for Skype. Skype subsequently informed eSafety that 'this only accounted for certain queues used by our content moderators, where re-review is required; not all queues of user reported concerns were accounted for.' Skype's amended response to the relevant question in the notice reported that the average time to respond from 24 January 2022 through 31 July 2022 was actually 2 days.

Skype provided additional information, stating that for the period July-December 2021, over 99% of content actioned for CSEA material on Microsoft's consumer services, including Skype, was detected through automated detection, rather than through user reports.

Skype also added that additional resources in the newly established Digital Trust and Safety engineering team of Microsoft (Skype's parent company) are working to improve automated detection of CSEA content.

## Questions about risk assessments when making policy or design changes to its services

**Conducting risk assessments**

Skype reported that it conducts safety risk assessments when it makes policy or design changes to its service and that all changes must comply with and be assessed against the Microsoft Digital Safety Standard (MDSS), which was launched in May 2022.

Skype also noted that Microsoft is working with the Digital Trust and Safety Partnership (DTSP), that Microsoft conducted a pilot of the DTSP Safe Framework in early 2022, and that Microsoft aims to roll the Framework out to all Microsoft services once the Safe Framework has been finalised. Skype added:

> Although Skype was not included in the initial Safe Assessment pilot, we anticipate more extensive risk assessment activities under the Microsoft Digital Safety Standard.

**Steps to mitigate risks identified and ensure safety by design**

When asked about the steps it takes to mitigate risks identified as part of its risk assessment process, Skype responded with the following:

> Services are reviewed against the Microsoft Digital Safety Standard MDSS, and the results are discussed with applicable senior leadership (CVP level) to align on remediation and resourcing.

Skype also stated that Microsoft used the DTSP Safe Assessment questionnaire, 'as a framework for examining the people, processes, and technology that contribute to managing content – and conduct-related risks', and pointed to the DTSP Safe Assessments Report for more information.

When asked how safety risks are evaluated alongside other types of risk, such as those identified in a privacy impact assessment, Skype responded that:

> Pursuant to the Microsoft Privacy Standard, privacy reviews are conducted to identify, assess, and mitigate potential privacy risks from the collection, storing and sharing of personal data when system capabilities or processes are being designed. Content moderation and other safety interventions often involve data processing and, as such, would be included in a data protection impact assessment.

Skype also highlighted that other risks may be 'service-specific' and require consultation with subject-matter experts in areas such as cybersecurity, lawful access, telecommunications, Responsible AI, accessibility, and corporate standards amongst others.

### Question about the use of metrics to internally assess efficacy of interventions to detect CSEA

Skype confirmed that metrics are in place and commented that a key Digital Safety Standard program objective was to use data to demonstrate how well content moderation systems are functioning, to ensure continuous improvement. Skype added that:

> Microsoft's Digital Safety Office, engineering and Responsible AI teams are working together to develop assessment methods for evaluating the measures currently in place to determine whether they are operating as intended and mitigating the intended risk(s).

# 6.6 Omegle Summary

## Overview

Omegle.com, LLC was only asked about its Omegle service.

### Omegle's responses

In September 2022, Omegle provided its initial response to a non-periodic notice that had been given to it by the eSafety Commissioner (initial response). At the time of the initial response, Omegle's Terms of Service permitted users between the ages of 13 and 18 years to use the service with parental permission and supervision.

eSafety posed some clarifying questions to Omegle. As part of its response to those questions in October 2022 (the subsequent response), Omegle advised eSafety that it had since changed its policies and now prohibits users under the age of 18 years from using all sections of the site. It also advised that the Omegle landing page, pop-up boxes and Terms of Service have been updated to reflect this change.

This change in policy means that several of Omegle's original responses may now be out of date. Responses that relate to the use of Omegle by persons aged under 13, and 13 to 17, are affected by this change.

For example, eSafety asked questions about age assurance and verification measures and what steps are in place to prevent children communicating with unknown adults. Omegle's response to these questions, and others, refer to services used and measures in place for the age groups under 13, and 13 to 17[40]. While Omegle has not provided further information on other changes made to steps taken, techniques and tools used, or resources in place to address this change in policy, Omegle pointed to its approach to age assurance in its initial response (a tick box confirmation of age – as summarised below in 'Questions about what age assurance or verification measures Omegle has in place). It indicated these measures and steps which previously applied to under 13s now also apply to those aged 13 to 17 since the change in policy.

**Note about Omegle Services**
Omegle has three main services – a moderated video chat service, an unmoderated video chat service and a text chat service.

[40]References to "ages 13 to 17" are inclusive of ages 13 and 17.

## Questions about what proactive measures Omegle has in place to minimise child sexual exploitation and abuse material

In its initial response, Omegle reported that it has a range of measures in place to proactively minimise CSEA material and activity, including:

- restrictions on use
- regular reminders on restrictions on use
- moderation (and consequent banning)
- proactive banning
- blocking certain 'common interest' topics
- working with law enforcement.

**Restrictions on use**

Omegle's initial response reported that it did not permit any part of its services to be used by children under the age of 13 years, and that children aged of 13 to 17 years were only permitted to use Omegle's moderated video chat or text chat service with the permission and under the supervision of a parent or legal guardian. Omegle also reported that children 13 to 17 years are not permitted to use the unmoderated video chat part of Omegle's service.

Omegle's subsequent response advised eSafety that it had changed its policies to prohibit children under the age of 18 years from using any of its services. It also advised that its landing page, age confirmation pop-up boxes and Terms of Service had been updated to reflect this change.

**Regular reminders on restrictions on use**

In its initial response, Omegle advised that users are required to acknowledge and represent that they are either 18 years or older or 13 years or older with parental supervision, for the moderated video chat and text chat, each time they try to access the Omegle service. For unmoderated video chat, users are required to acknowledge and represent each time that they are 18 years or older, or the age of legal majority in their jurisdiction.

eSafety's own understanding is that as a result of Omegle's change to its Terms of Use, which now prohibit users under the age of 18 years from accessing its services, users must now confirm via a check-box that they are 18 years or older before using any of its services.

**Moderation**

Omegle provided the following information regarding steps on its 'moderated video chat' service:

- Omegle stated that automatic capturing of screenshots takes place, with review using Hive AI. Omegle noted that where 'Hive's AI has high confidence [that images] are sexual in nature, an immediate and automatic ban is issued on the user's use of the moderated section of Omegle's chat service'.
- Where Hive's AI detects 'a significant chance (contrasted with high confidence) of containing either sexual or suggestive content, but not with a sufficiently high degree of confidence, a ban is not issued from the AI's decision alone, but is passed on to human review.' Omegle reported that moderators can issue varying lengths and types of bans and reports to appropriate authorities based on this flagging.
- Moderation is contracted to Gracall International, which provides human content moderation services to Omegle. Omegle stated there were at least three human moderators 'at any given time' and it is currently trialling having four moderators working for 24 hours one day out of each week.
- Omegle also stated that there is automatic flagging of messages that contain unique phrases which are known to be associated with CSEA on the moderated video chat service and text chat service.
- Omegle did not provide information on proactive measures implemented on the unmoderated video chat.

Omegle provided the following information regarding the moderation of the text chat service:

- Automated scanning using a program run by Omegle for certain patterns of text on their text chat service.
- Text chats are also passed on for human review. Human moderation of the text chat service is conducted by one Omegle employee who can ban users and report content to NCMEC.

**Proactive banning**

Omegle stated that it also takes steps to proactively ban certain users. eSafety has decided not to publish further details.

**Blocking common interest topics**

The 'common interest' function allows users to specify an interest and connect with other users whoshare that same interest. Omegle stated that it prevents the use of 'common interest' topics or tags which have the potential to be used in association with CSEA. Omegle stated that a list of common interest topics and tags uses a variety of sources.

**Working with law enforcement**

Omegle stated that it works with global law enforcement agencies in relation to CSEA material and criminal activity more generally. Omegle stated that this includes reporting material to NCMEC, assisting with law enforcement requests to access certain information about users, and providing information to law enforcement where a user has been identified as engaging in harmful or inappropriate conduct.

### Questions about what measures Omegle has in place to prevent children communicating with adults they do not know on Omegle

In its initial response, Omegle stated that it does not prevent children communicating with adult users they do not know. Omegle added that users are asked when entering the service to acknowledge that they are over the age of 18 years, or over the age of 13 years with the permission and supervision of their legal guardian.

While Omegle did not provide any additional information in its subsequent response to eSafety about whether steps are being (or will be) taken to remove any users under the age of 18 years, to align with its recent policy change, Omegle pointed to its approach to age assurance in its initial response (tick box confirmation – as summarised below) which also applies to under 18s.

Omegle noted that the anonymity of users to one another is intended to protect the privacy of users. Omegle stated that users are not anonymous to Omegle, and although Omegle does not require personal information to be provided to use its service, it collects other information to be able to take appropriate steps to ban users who may be engaging in unlawful or harmful conduct.

### Questions about what age assurance or verification measures Omegle has in place

In its initial response, Omegle stated its terms only permit users aged 13 to 17 years of age to use its moderated chat services with the permission and under the supervision of a parent or legal guardian. To enforce this rule, Omegle requires the user to tick a box agreeing to those terms. If content is flagged by AI tools, human moderators will ban a user suspected to be under 13 years.

Omegle also stated in its initial response that users aged 13 to 17 years are not permitted to use Omegle's 'unmoderated video chat' service. eSafety notes that users are required to tick a box to confirm their age as 18 or older on entry to the 'unmoderated video chat' service.

Omegle commented that age verification is difficult to enforce without linking to some other form of system (such as a bank account) or requesting a form of identification, which it commented raises issues including useability and privacy. Omegle advised that it takes age verification seriously and keeps abreast of age verification technologies and industry standards around the globe. Omegle stated:

> For example, Omegle is aware of and following the consultation process for the eSafety Commissioner's Age Verification Roadmap. While we appreciate the Age Verification Roadmap relates to online pornography, we expect the information gathered during the consultation process may assist other online services like Omegle to enhance age verification and ensure we are keeping up with best practice.

In its subsequent response that identified the change to its policy prohibiting users under the age of 18 years from using all sections of its site, Omegle did not provide additional information about additional age assurance or verification measures Omegle has in place to enforce the policy change, though Omegle pointed to its initial response in relation to age assurance.

### Questions about any techniques or language analysis used by Omegle to detect potential online grooming and whether Omegle allows the sharing of personal details and other social media contact details on the service

Omegle stated that it has an automated scanning system which monitors text chats for certain patterns of text including grooming. Omegle stated that it can add to the dataset of text patterns so that similar patterns can be detected again in the future.

Omegle suggested that the risk of children being exposed to grooming over a long period of time on Omegle is low, because chats only last for as long as both parties keep the URL open. Once a user exits the conversation users can't reconnect with each other on Omegle.

Omegle stated that it recognises there are risks if children are redirected to other social media sites during an Omegle chat. Omegle stated this is one of the reasons it proactively monitors for specific information which is known to be in association with CSEA material and discourage users from disclosing personal information during chats.

Omegle reported it does not prohibit or prevent users from sharing their own personal details, including other social media contact details, when using the service, although this is discouraged. Omegle stated that unless a user shares their own personal details by their own discretion, personal details are not available to other users. Users in the chat function appear to other users as 'you' and 'stranger'.

### Questions about whether users can make reports to Omegle about other users or specific chats for breach of Omegle's terms of service, including for CSEA activity

Omegle noted that "the Community Guidelines state that 'If you encounter conduct or content on the Services that you believe violate any of the guidelines below or if you have other safety concerns about the Services, please contact Omegle at **safety@omegle.com** and include 'Omegle-Safety' in the subject line.'"

Omegle reported that users' text chat logs can be uploaded to Omegle's servers at the users' discretion. If no chat log is uploaded, Omegle advised that it can trace a reported user if the reporter provides the time of the relevant chat.

### Question about what measures are in place to prevent Omegle users who are suspended or banned for CSEA-related breaches from using the service

Omegle identified interventions that it has put in place to prevent recidivism, which rely on a minimal number of different indicators. eSafety has chosen not to publish the number or type of identifiers used. Omegle added that if any CSEA material is detected from a user, whether via moderation or via a report from another user, the user will be permanently banned from using Omegle and reported to appropriate authorities.

Omegle advised that moderation includes processes to automatically flag messages that contain unique phrases that are associated with abuse by a particular person.

### Questions about risk assessments when making policy or design changes

**Conducting risk assessments**
Omegle reported that it conducts safety risk assessments when making policy or design changes to its services.

Omegle referred to a data protection impact assessment (DPIA) conducted in April 2022 in connection with the European Union's General Data Protection Regulation and advised that although the DPIA focusses on data protection, it also considered other risks to children including CSEA and access to harmful or inappropriate content.

**Steps to mitigate risks identified and ensure safety by design**
Omegle reported that it takes steps to mitigate safety risks and provided the following examples:

- 'Spy mode' was previously a feature on Omegle that allowed a three-person text chat, where one person would propose a prompt for the other two to discuss. Omegle removed this feature determining it was too difficult to effectively moderate.
- Omegle reported and provided details about steps taken to detect ban evasion. eSafety has decided not to publish the details.
- Omegle reported that it has recently strengthened its Terms of Service, Privacy Policy and Community Guidelines.

In its subsequent response, Omegle also referred to its recent policy change to prohibit users under the age of 18 years from using all sections of Omegle's site as 'an example of Omegle's ongoing evaluation of, and modification to, its service based on a variety of considerations, including for example technology changes and safety assessments'.

When asked how safety risks are evaluated alongside other types of risk, such as those identified in a privacy impact assessment, Omegle responded that it evaluates safety risks both alongside its Data Protection Impact Assessment and independently , through the day-to-day running of its business including by:

- considering safety risks when a policy or design change is made
- making design changes to address new information about safety risks.

### Question about the use of metrics to internally assess efficacy of interventions to detect CSEA

Omegle reported it has various metrics in place as part of its moderation system. Omegle stated that, with regards to human monitoring of moderated video chat by an external contractor, it has the following metrics in place:

- metric to ensure that 100% of images flagged by AI for human review are in fact reviewed by humans (in other words, that staffing is adequate)
- number of snapshots banned
- ban percentage (number of bans divided by the number reviewed)
- time spent working (minutes per day)
- bans per work-hour (bans divided by total hours of work in a given day).

Omegle stated that it investigates any anomaly raised by the statistics, for example a particularly low ban rate for a specific month.

After engagement on follow-up questions, Omegle advised that human review of messages from Omegle text chats is conducted in-house by one Omegle staff member. Omegle also stated that it uses the number of reports made to NCMEC as a metric to assess the success of its interventions to detect.

Omegle also stated that the AI moderation system has built-in checks and the threshold for images is set broadly enough that when the AI detects a low likelihood of sexual content it will still go to a human moderator for review.

# 6.7 Snap summary

## Overview

Snap Inc. was asked only asked about its Snapchat service.

### Questions about steps taken to detect known and new child sexual exploitation and abuse material

In relation to questions about hash matching for known CSEA images and videos on parts of its service, Snap confirmed that it uses the hash matching tools PhotoDNA for images and CSAI Match for video on the following services.

| Hash matching tools for known CSEA images and video are used | Hash matching tools for known CSEA images and video are not used |
|---|---|
| Discover | Snaps |
| Spotlight | |
| Direct Chat (chat media) | |
| Note: Snap stated it **scans images and video uploaded from a user's phone to Snapchat** on Snap map, Chat, Discover and Spotlight.<br>Snap noted that it would not be a realistic possibility to match Snaps to known CSEA hashes because the Snaps are taken in real time. | |

Snap referred to PhotoDNA as 'a proprietary image-identification and content filtering technology widely used by online service providers to detect known CSEA in still images. It is widely acknowledged as the de facto industry standard for known CSEA detection in images.'

In response to a question about other reasonable steps being taken to proactively minimise CSEA material, Snap described its approach as being broader than reactive content moderation. Snap commented that this involves a focus on prevention and safety by design. Snap reported the following:

- It does not offer an open news feed that gives the opportunity to broadcast illegal content such as CSEA material.
- There are protections and features built into Snapchat's original and current design that prevent people from publicly broadcasting content without it first being moderated, and that limits strangers contacting people they don't know.
- The public areas of Snapchat are the Discover page for news and entertainment and the Spotlight tab for 'the community's best Snaps'. It said these have measures to ensure harmful or illegal content is not surfaced to large numbers of people, stating that 'our Discover page for news and entertainment, and our Spotlight tab for the community's best Snaps – are substantially curated and moderated, respectively.'

- There are systems and processes in place, which Snap describes as effective, and which Snap notes enables it to act quickly when it learns of illegal and harmful content and activity. These processes include easy-to-use in-app reporting tools and safety teams working 24/7 to review user reports and take appropriate action.
- Safety teams may take various actions including warning an account, deleting content in question, terminating an account and deleting a user's data, and/or reporting to law enforcement in certain circumstances. In the case of a report of CSEA material on an account, Snap advises that the likely outcome would be terminating the account and making a report to the National Center for Missing and Exploited Children (NCMEC).
- Snap was committed to working with other service providers and non-government organisations to improve online safety and tackle online child exploitation. It identified that it:
  - co-chairs the engineering working group for the cross-industry online child safety organisation, the Tech Coalition and chairs its Collective Action working group
  - hosted Tech Coalition's first child safety hackathon in September 2022
  - conducts and contributes to research projects and other programs aimed at developing technical solutions for all companies.

**Detecting new CSEA material**

Snap reported that it does not use any automated tools to detect new CSEA material on the following parts of its service:

- Discover
- Spotlight
- Direct Chat
- Snaps

Snap noted the following:

> We are not aware of technology that is widely available and suitably reliable to detect indicia of so-called new or 'first-generation' CSEA material to sufficiently high levels of accuracy. Given the privacy implications of monitoring or scanning users' communications, the threshold for the application of such technology is necessarily very high. Snap is exploring the development and use of technologies that could permit detection of new or un-hashed CSEA material.

In providing additional context and information on any alternative and reasonable steps to proactively minimise the provision of new CSEA material on those parts of Snapchat, Snap reiterated the following measures (which it also reported as being used for known CSEA material):

- curating and moderating the public areas of Snapchat
- providing in-app reporting tools and review by its safety teams
- reporting CSEA material to NCMEC.

### Questions about steps to identify whether an account that has previously shared CSEA material contains other CSEA material

Snap stated that when CSEA material is identified, in 'certain instances when additional evidence is required', it uses technical tools and moderator review to identify whether other content shared by the account contains CSEA.

In response to a further question about the circumstances in which a review of other material on an account occurs, Snap subsequently clarified that:

> Review of account contents may be required if, for example, Snap receives a report of CSEA activity by a particular Snapchat account, but the report does not identify any particular piece of content as CSEA content. In such instances, Snap's Trust & Safety moderators will review accounts' contents to confirm or deny the reports and determine whether reporting to NCMEC is required (i.e., whether there is evidence of CSEA).
>
> If Trust & Safety moderators identify evidence of CSEA, the moderators will report the account(s) to NCMEC and may cease further review, in order to avoid unnecessarily exposing the moderator to additional, harmful CSEA content.

### Questions about any language analysis technology used to detect potential online grooming conversations

Snap stated it does not use any language analysis technology to detect potential grooming conversations on any parts of its service.

In response to a question about alternative steps being taken to proactively minimise the grooming of children on Snapchat, Snap stated that it has made a number of design choices to seek to proactively minimise the risk of grooming of children through its 'safety by design framework', including the following:

- Snap reported that it was harder to find others on Snapchat compared to other platforms – for example, Snap reported that friends lists are only visible to the individual user (except where the new parental and caregiver tool Family Centre is in use).
- Snap reported that default settings on Snapchat are such that a user cannot receive a message from anyone who they have not accepted as a friend on the app, or who is not a contact in their phone book. As noted below, Snap subsequently confirmed that messages could be 'sent' to users under 18 who are not friends, but by default, users cannot read these messages unless they accept the sender as a friend.
- Snap highlighted that location sharing on Snap Map is off by default ('Ghost Mode'), and there is no option on Snapchat to share location with anyone other than a user's friends or a designated sub-set of those friends.
- Snap commented that users have the ability to report content they find concerning through 'simple, intuitive and easy-to-use in-app reporting tools'. Snap noted that there is a dedicated channel on the Discover page called Safety Snapshot that provides advice for users on keeping their account secure.

- Snap noted that the Family Centre gives parents the ability to know 'who their teens are friends with on Snapchat and who they have been communicating with over the past 7 days, without revealing the substance of their teen's conversations.' Parents and carers have the ability to report suspicious accounts to Snap for review. The 'Friend Check Up' feature was also highlighted as a means to prompt Snapchat users to review who they are friends with and make sure the list is made up of people they know and still want to be connected with.

Snap stated these measures:

> work to proactively minimise the risk of harmful contact from strangers towards Snapchatters of all ages. They make it difficult for strangers to identify, much less meet, younger users:
> potential predators cannot browse for people to target or view other users' personal information such as their age, school, interests or location. It's never possible to see the location of someone who isn't a friend.

Snap added that in relation to users' private communications, being private chat messages and Snaps, 'our community has a justifiable expectation that we are not monitoring their every communication, and that is not something we do.'

Snap also stated that:

> We are not aware of technology that is widely available and suitably reliable to detect indicia of child grooming for sexual purposes in text conversation to sufficiently high levels of accuracy. Given the privacy implications of monitoring or scanning users' private communications, the threshold for the application of such technology is necessarily very high.

### Questions about any age assurance or verification measures Snap has in place and use of indicators to assess whether a user may have provided a false age on sign-up

Snap responded that Snapchat is designed to appeal to teen and adult audiences and that individuals under the age of 13 years are not permitted to create Snapchat accounts.

Snap responded it does not market Snapchat to children and Snapchat is not available in the 'Kids' or 'Family' sections of any app store; the app is rated 12+ in the Apple app store and rated Teen in the Google Play Store.

Snap stated it has the following measures in place to ensure age-appropriate use on Snap:

- Registration requires a date of birth. Snap reports that registration fails if a user is under the age of 13 years.
- If Snap is made aware that a Snapchat user is under the age of 13 years by a user, a parent or law enforcement report, Snap terminates the account and deletes the user's data. Snap can also take other measures which eSafety has chosen not to publish.
- Snap prevents a user ages 13 to 17 years old from updating their year of birth to an age over 18 years.

Snap added that it considers age assurance can be disproportionately intrusive and against what it described as its core privacy-protecting principles on minimising the type of data it collects. Snap noted that age verification often requires access to, collection and retention of identity documents and expressed a view that it may introduce bias and inaccuracy where techniques rely on profiling or AI facial analysis. Snap noted that there is a risk that some age assurance methods may result in exclusion or discrimination of already marginalised groups due to 'bias, inaccuracy or requirements for official documentation.'

Snap added that:

> Given the strong protections that we have in place to protect Snapchatters of all ages – as set out throughout this response – we consider our approach to age assurance on Snapchat appropriate and proportionate. Age assurance of children is complex and there are several data protection and technological challenges which remain unresolved at this time. We are committed to developing robust approaches to age assurance, but we do not do so at the detriment of other safeguards integral to preserving the safety and trust of our community.

### Questions about steps and indicators used to prevent recidivism

Snap identified interventions that it has put in place to prevent recidivism, which rely on several different indicators. eSafety has chosen not to publish the specific number or type of indicators used.

Snap stated that additional indicators are used 'when doing deeper investigations.'

### Questions about privacy settings for children

Snap confirmed that children are given the most restrictive privacy settings by default, but that children can change their contact settings to a less restrictive option. At the time that Snap responded to the notices, this included an 'everyone' setting which allows anyone who has your Snapchat username to contact you.

In answer to a question about the percentage of children who change their privacy settings to a less restrictive option, Snap stated that approximately 6% of users aged 13 to 17 in Australia had changed their contact settings to 'everyone'.

In November 2022, Snap subsequently notified eSafety of a change to its contact settings, removing the 'everyone' option for users aged 13 to 17, and replacing it with a 'My Friends and Their Friends' option. Snap reported that this change means that users will need to have at least one mutual friend in common to be able to receive a message. Snap noted that it was still updating the visual settings for this change, but that 'in practice the change has already been made – users under the age of 18 now need to have at least one friend in common to receive a message from someone who they have not accepted as a friend.'

Snap added that the 'Friends and Contacts' setting (which allows only a user's Snapchat friends and people they have added to Contacts to contact them) is currently being tested and is not available to users under the age of 18.

Snap advised that, by default, users under the age of 18 cannot receive a message from someone they have not accepted as a friend. eSafety asked further questions about this, summarised in the following section.

Snap advised that location sharing on Snap Map is off by default for all users and that there is no option on Snap Map to share one's location with anyone other than friends, or a designated sub-set of those friends.

The Public Profile feature (which allows users to create a profile that is viewable by non-friends) is restricted to users 18 and older.

### Questions about messages from unknown users

Snap was asked whether users can send messages or photos from users that have not already been added and accepted by them as 'friends'.

Snap responded that the default settings for all users under the age of 18 years only allow a user to receive messages or photos from another user if they have accepted the other user as a friend.

In response to a follow-up question from eSafety, Snap confirmed that messages can be sent to users who have not already been added and accepted as friends. eSafety's understanding is that these messages and photos can then be read when a friend is accepted.

Snap outlined that there are additional settings in place to make it harder for teens to be contacted by people they don't know, including the following:

- Teens can only show up as a suggested friend to another user in limited instances, for example, if they have multiple friends in common.
- Users under the age of 18 years can create content which is viewable by non-friends on Spotlight but they cannot create 'persistent, attributed content that can be browsed by non-friends'. The lack of attribution means that viewers of that content cannot navigate back to the user's profile and send them a friend request.

### Questions about user reporting

Snap was asked about the average time to respond to user reports of CSEA content. Snap provided the following information.

| **Median time to action user reports of CSEA (e.g. content removed and reported, user banned, or other content moderation decision taken)** | |
|---|---|
| Snap | 4 mins |

## Question about how CSEA is detected

Snap was asked for the percentage of removed CSEA material on Snapchat identified proactively by automated tools (vs material reported by users, trusted flaggers, or identified by moderators).

| CSEA content identified by | Proportion |
|---|---|
| Automated tools | 87.5% |
| User report | 12% |
| Trusted flagger | 0.5% |
| Moderators | n/a |
| Other | n/a |

Snap added that:

> Account deletions resulting from trusted flagger reports comprised approximately 0.5% of total account deletions for CSEA material reported in our most recent Transparency Report, or 1,058 out of a total 198,109 enforcements.

Snap stated the value for Moderators is 'N/A', as this information is provided in the categories in the table in the previous section, noting:

> Our moderators review all user reports to determine whether they contain CSEA. Moreover, they review tasks prompted from our automated tools' detections and enforcements to help ensure the accuracy of, and further refine, these tools.

## Questions about risk assessments when Snap makes policy or design changes to its services

**Conducting risk assessments**

Snap reported it conducts safety risk assessments when making policy or design changes to its service.

Snap stated its safety by design and privacy by design practices are embedded in risk assessment of any new product or feature. Snap stated that its Legal, Policy and Trust and Safety teams review products from inception to post-launch, considering any potential impacts on users' safety and privacy.

Snap also stated that it involves external experts at key points in the development process and provided the example of the development of their Family Centre tool.

Snap stated these design principles and review 'are reflected in the design and build of Snapchat, and in practical features that enhance user safety and protect their privacy.'

Snap added that:

> Our teams continually assess the effectiveness of our systems and work to improve them to help ensure the safety of our Snapchat community. For example, we track all reports received and actions taken, and we have teams dedicated to increasing proactive detection of abuse. Policies are subject to continuous review, reflecting the constantly evolving nature of harmful and illegal activity online.

Snap advised that internal-facing policies are updated on a regular basis.

**Steps to mitigate risks identified and ensure safety by design**

Snap responded that safety by design review is a 'crucial part of the design and development process' and provided the following examples of this approach in action:

- Spotlight is a recent feature which allows users to submit their best snaps for viewing by the wider community. Spotlight Snaps are unattributable by default for users under the age of 18 years, which means that the wider community cannot contact that user or look at their profile or content.
- Snap Map is set to 'Ghost Mode' by default which means that location sharing is off. Snap reported that the use of the ghost icon and accompanying tutorial for Snap Map is to ensure that this setting is easily understandable to younger users. Snap Map also regularly reminds users who have opted to share their location, that they are doing so.

When asked how safety risks are evaluated alongside other types of risk, such as those identified in a privacy impact assessment, Snap responded:

> We consider both safety by design and privacy by design to be fundamental principles and processes at Snap, and these processes go hand in hand. Indeed, we find that measures designed to protect users' privacy often also enhance their safety, and vice versa.

### Question about the use of metrics to internally assess efficacy of interventions to detect CSEA

Snap responded that it does have metrics in place to internally assess the efficacy of its interventions to detect CSEA and pointed to their bi-annual transparency reports, commenting that it is the only major platform to provide country-specific breakdowns of content reported and enforced, including a dedicated page for Australia.

Snap stated that, through the Tech Coalition, it is involved in a new effort, being led by Childlight, to agree cross-sector metrics for both the detection and prevention of CSEA.

esafety.gov.au