



Basic Online Safety Expectations

Summary of industry responses to
mandatory transparency notices

October 2023

Contents

1. Executive Summary	3
2. Glossary	7
3. Key Findings	10
4. The Basic Online Safety Expectations	19
5. Information about the Notices.....	19
5.1 Who received Notices?.....	19
5.2 What questions did eSafety ask?	21
5.3 What was the Notice process?	23
5.4 What process was followed once the information was received?.....	23
5.5 What information has been published, and what has been excluded?.....	24
5.6 What happens next?	25
6. Non-compliance with the Notices and action taken by eSafety	26
6.1 eSafety’s powers to require reports on compliance with the Basic Online Safety Expectations	26
6.2 Findings of non-compliance	26
6.3 Why it is important that providers comply with reporting Notices	28
7. Transparency: Responses by Issue	29
7.1 Detecting previously confirmed (‘known’) images.....	29
7.2 Detecting previously confirmed (‘known’) videos	31
7.3 Blocking links (URLs) to known CSEA material.....	33
7.4 Detecting new material.....	34
7.5 Proactive detection	37
7.6 Detecting CSEA in livestreams or video calls	38
7.7 Detecting Grooming.....	40
7.8 Detecting sexual extortion and other exploitation of children	43
7.9 Detecting underage users.....	46
7.10 Detecting recidivism.....	49
7.11 Recommender System safety	52
7.12 User Reporting.....	56
7.13 Languages providers operate across	60
7.14 Roles and responsibilities of volunteer moderators	62

8. Transparency summaries: Individual provider responses 64

8.1 Google Summary..... 64

8.2 Twitter Summary 92

8.3 TikTok Summary 114

8.4 Twitch Summary.....135

8.5 Discord Summary155

1. Executive Summary

On 22 February 2023 the eSafety Commissioner (**eSafety**) issued a second set of non-periodic reporting notices (the **Notices**) pursuant to section 56(2) of the Online Safety Act 2021 (the **Act**). The notices were issued in relation to Google, Twitter (which later became known as X)¹, TikTok, Discord and Twitch. The [Basic Online Safety Expectations](#) are set by the Australian Government and provided for by the Act.

The Notices focus on understanding the steps being taken to address child sexual exploitation and abuse, as a particularly high risk and high harm issue that has seen sustained growth.

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

In addition to these questions, providers were also asked about the tools, policies and processes they are using to address the sexual extortion of children and to avoid the risk of amplifying harmful content through recommender systems.

These new Notices build on the findings from eSafety's [first transparency report](#) published in December 2022, which compelled transparency on the steps Apple, Microsoft, Skype, Meta, WhatsApp, Snap, and Omegle, were and were not, taking to detect and address child sexual exploitation and abuse, following notices issued to these companies in August 2022.

When issuing the notices in February 2023, the Commissioner noted that:

Our first set of notices sent to companies including Apple, Meta and Microsoft in August last year, revealed many companies were not taking even relatively simple steps to protect children and are failing to use widely available PhotoDNA technology to detect and remove child abuse material.

No solution ever presents itself by ignoring the problem. We need the companies to start turning the lights on, so we can get a true sense of the size and scope of this problem.

Implementation of tools matters as well as the existence of them.

It's time for all members of the online industry to step up and use their financial, intellectual, and technical resources to identify and remove this material from their platforms because even one child sexual exploitation image, is one too many.

¹ In July 2023 the Twitter service was rebranded as 'X', however it was known as Twitter at all points during the notice period, and so is referred to as Twitter in this report. The notice was given in February 2023 to Twitter, Inc. which subsequently merged into X Corp.

Non-compliance with the notices

eSafety closely considered the responses provided by the five recipients of the Notices. eSafety found that Google and Twitter did not comply with the Notices given to them to the extent that they were able.

Google provided answers in certain instances that were not relevant, or were generic, and in other instances provided aggregated information across multiple services where information regarding specific services was required. Google has been given a formal warning, notifying the company of its failure to comply and warning against non-compliance in the future.

eSafety considered Twitter's failure to comply to be more serious. In some instances Twitter failed to provide any response to the question, such as by leaving the boxes entirely blank. In other instances, Twitter provided a response that was otherwise incomplete and/or inaccurate. Despite Twitter being given further opportunities to provide the information required by the Notice, significant gaps remain. eSafety has given a service provider notification to Twitter, notifying it of the non-compliance. The service provider notification has also been published on eSafety's website. Twitter has also been given an infringement notice for AUD\$610,500 for its non-compliance. Twitter has 28 days to request the withdrawal of the infringement notice or to pay the penalty. If Twitter chooses not to pay the infringement notice, it is open to the Commissioner to take other action.

Details of the actions taken for this non-compliance are outlined in section 6. eSafety will continue to use the full range of powers available to eSafety to ensure transparency and to hold providers to account.

Significant variation in safety steps

In a continuing theme to the December 2022 report, this report again shows significant variation in the steps being taken by providers to protect users and the wider Australian public. Similar to the 2022 report, eSafety discovered that there is no common baseline in terms of the safety protections in place. Detection tools are not being used consistently across different services, even where multiple services are owned by the same parent company. Key findings of the Notices include:

- Discord, and Google services covered by the Notice², do not block URLs linking to known CSEA³, despite the availability of verified lists of CSEA URLs being made available by expert organisations such as the Internet Watch Foundation (IWF). Google confirmed it uses URL lists on its search service, although this was not in scope of the Notice⁴.
- Google stated it is using tools to detect CSEA material stored and shared in Google Drive, demonstrating the technical feasibility of this important intervention. In comparison, the December 2022 report found that Microsoft and Apple were not scanning for CSEA material stored in OneDrive (Microsoft only uses tools when material is shared) and that Apple was not scanning for this material on iCloud either when stored or shared.
- Despite making its technology – CSAI Match – available to other services Google does not use it to detect known CSEA videos on Gmail, Messages and Chat. Discord, Twitch and Twitter also do not use any tools to detect known CSEA videos on their services. TikTok does use video hash matching tools.
- For the three months after Twitter’s change in ownership, the proactive detection of CSEA fell from 90% to 75%. Since then Twitter stated that its proactive detection has improved.
- TikTok and Twitch use language analysis technology to detect grooming of children in public parts of their services and in direct messages, whereas Twitter and Discord do not use tools to detect grooming on any parts of their services.
- All five providers had livestreaming features on their services during the Report Period, but only Google, TikTok and Twitch stated that they had measures in place to detect livestreamed CSEA. Discord stated it does not have any measures in place to detect CSEA in livestreams. Twitter’s response did not provide the information required.⁵
- There are significant differences between services in terms of the languages covered by their human moderators. Google and TikTok have human moderators operating in more than 70 languages, whilst Twitter moderators cover 12, Discord cover 29, and Twitch 24. Some of the most widely spoken languages in Australia are not covered by default by Twitter, Discord and Twitch.
- For Discord and Twitch, who are partly community moderated, professional safety staff are not automatically notified when a volunteer moderator identifies child sexual exploitation and abuse material.

² See page 19 for the list of services.

³ 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.

⁴ Under the Act the basic online safety expectations apply to social media services, relevant electronic services and designated internet services that can impact the online safety of Australians. This includes social media, messaging, gaming, dating, file sharing services and other apps and websites but does not extend to search services.

⁵ Following publication of this Transparency Report X Corp. advised eSafety that its response to questions regarding the detection of livestreaming of CSEA on the Twitter service ‘was inaccurate due to an inadvertent error’ and provided a revised response to the question. X Corp.’s revised response is [available](#)

- The time providers take to respond to user reports of child sexual exploitation varies significantly across services. The median time to respond varies from 5.2 minutes for material shared publicly on TikTok to 13 hours for direct messages on Discord. This compares with the 2022 transparency report, which found that Snap responded to user reports within 4 minutes, whereas Microsoft's Skype, Teams, and OneDrive took a median time of 2 days, and 19 days where cases needed 're-review'.
- Google (Gmail and Google Messages), Twitter (Direct Messages) and Discord (Livestreams) have no in-service reporting options. Users are required to contact the provider via webform if they wish to complain about illegal or harmful activity. Twitch requires users to be signed into accounts before they can make an in-service report.
- Measures to detect recidivism⁶ vary across providers. Discord, TikTok, Twitch and Twitter check multiple indicators to try and detect repeat abusers. Google services only check a minimal number of indicators.

Furthering transparency

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 as effective or appropriate on another.

However, where the Notices have provided information of potential failure to implement the Expectations, eSafety will be engaging with those providers to understand their plans to address these safety shortcomings and any obstacles to compliance.

As of the date of this report, eSafety has issued 13 Notices⁷ to providers. Although this constitutes a small number, the Notices have gone to providers of some of the most widely used online services in Australia and globally. The information received and published in this report, together with the December 2022 report, represents a significant step towards greater transparency and understanding of what providers are and are not doing to protect Australians online.

⁶ In the context of this report, banned or suspended users re-registering to an online service with new details to continue perpetrating online abuse. This can take the form of multiple or fake imposter accounts, including by automated accounts.

⁷ A section 56(2) non-periodic reporting notice was issued to X Corp. (Twitter) on 22 May 2023, focussing on the steps that Twitter is taking to address online hate on its platform.

2. Glossary

- **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 tools, in the context of this report, to support the moderation of online content.
- **Automated tools:** 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:** Individuals 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, as well as other types of material.
- **CSEA:** child sexual exploitation and abuse. CSEA encompasses both 'child sexual exploitation' (a broad category of content that encompasses material and activity that sexualises and is exploitative to the child, but that does not necessarily involve the child's sexual abuse) and 'child sexual abuse' (which involves sexual assault against a child). Child sexual abuse is a narrower category and can be considered a sub-set of child sexual exploitation.
- **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 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 certain behaviour such as grooming. May also refer to use of keywords.
- **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.
- **Machine Learning:** The patterns derived from training data using machine learning algorithms, which can be applied to new data for prediction or decision-making purposes.
- **Machine Learning Classifiers:** A classifier is an algorithm that automatically orders or categorizes data into one or more of a set of ‘classes.’ They are rules that map input data into predefined categories.
- **Material:** Defined in section 5 of the Act to mean material, whether in the form of text, data, speech, music or other sounds, visual images (moving or otherwise), any other form or any combination of forms. In the context of CSEA, ‘material’ typically refers primarily to images or video content.
- **Natural Language Processing:** Also known as NLP, refers to the branch of computer science—and more specifically, the branch of artificial intelligence or AI—concerned with giving computers the ability, through algorithms, to understand text and spoken words in much the same way human beings can.

- **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 that has not been previously confirmed to contain content depicting child sexual exploitation or abuse, and which has not previously been hashed and stored in a hash database. Also known as, ‘first generation CSEA material’.
- **Notice:** For the purpose of this report, a Notice is a non-periodic notice given to a provider under section 56(2) of the Act on 22 February 2023.
- **Recidivism:** In the context of this report, banned or suspended users re-registering to an online service with new details to continue perpetrating online abuse. This can take the form of multiple or fake imposter accounts, including by automated accounts.
- **Recommender Systems:** Also known as content curation systems, are the systems that prioritise content or make personalised content suggestions to users of online services.
- **Report Period:** When online service providers receive a reporting Notice from eSafety they are required to prepare a report about the extent to which they complied with the basic online safety expectations during a specified period. This period is referred to as the Report Period. The Report Period for this set of Notices is 24 January 2022 to 31 January 2023. Information provided should reflect this period, unless stated otherwise.
- **Sexual extortion:** Also known as sextortion, is a crime involving online blackmail, where victims are tricked into sending intimate images of themselves to someone who then threatens to share the images unless demands are met, usually for payment. There has been a substantial growth in sexual extortion, targeting teenage males in particular. In some cases, depending on the age of the victim, this amounts to child sexual exploitation.
- **The Act:** The Online Safety Act 2021 (Cth).
- **Trusted Flagger:** An individual or entity which is considered to have particular expertise and responsibilities for the purposes of tackling harmful content online.

- **Underage:** Under 13 years old on most social media and gaming platforms. The US Children’s Online Privacy Protection Act (COPPA) of 1998 states 13 as the minimum age requirement and this age has largely become the de facto age standard for many social media and other services online.
- **URL database:** Databases of URLs linking to known CSEA material are kept to enable providers to prevent their use on a service. An example of a URL database is the one provided by the Internet Watch Foundation.⁸
- **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. Key Findings

The following pages outline some of the key findings from the Notices including steps taken by services to detect livestreamed CSEA, grooming of children, the blocking of URLs to known CSEA, community moderation methods, languages moderators operate across, and user reporting.

These key findings are accompanied by quotes from third party and eSafety insights on the nature of these harms and are also set alongside the number of CSEA reports made by service providers to the National Center for Missing and Exploited Children’s (NCMEC) Cybertipline.

⁸ [Internet Watch Foundation URL List](#), accessed 9 October 2023


Detecting livestreamed child sexual exploitation and abuse

Livestreamed CSEA involves the broadcasting of acts of sexual exploitation or abuse of children via webcam or video to people anywhere in the world. The sexual predator controlling the child sometimes charges money for providing access to the livestream.

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

Use of technology to detect child sexual abuse livestreams*

Not using:

 **Discord** (public and private servers, direct messages)

Using:

 **Google (YouTube)**

 **TikTok**

 **Twitch**



Discord does not monitor or record livestream content or voice chats. Discord has prioritized resources into other forms of CSEA detection, as running models across this type of content at the scale Discord operates would be prohibitively expensive and would operate at the detriment of other Discord safety programs.'

Discord response to the notice question asking about measures in place to prevent the livestreaming of child sexual exploitation and abuse on Discord.

*Following publication of this Key findings report X Corp. advised eSafety that its response to questions regarding the detection of livestreaming of CSEA on the Twitter service 'was inaccurate due to an inadvertent error' and provided a revised response to the question. X Corp.'s revised response is [available](#)

eSafety and third party insights into CSEA harms



This form of child sexual abuse online, along with "self-generated" abuse in livestreams, are all live crime scenes... committed daily on online platforms.'

International Justice Mission (IJM) 2022
[IJM Submission to Public Consultation on the Draft Consolidated Industry Codes of Practice for the Online Industry \(Class 1A and 1B Material\)](#)



Australian children as young as eight are being coerced into performing live-streamed sexual acts by online predators, who often record and share the videos on the dark net and sexually extort victims into producing even more graphic content.'




Australian Federal Police (AFP) 2021
[AFP warn about fast growing online child abuse trend | Australian Federal Police](#)

Detecting grooming




eSafety is aware that sexual predators use online services to 'groom' children. Grooming is predatory conduct to prepare a child or young person for sexual activity, so they can be tricked into sending images or videos or behaving sexually in video livestreams. In addition, the child may be sexually extorted into sending more material or money if the predator threatens to share the image or video.

Use of language analysis technology to detect likely online grooming

Not using:

-  **Discord** (public and private servers, direct messages)
-  **Google** (Meet, Chat, Messages, Gmail)
-  **Twitter** (Tweets, direct messages)

Using:

-  **Google (YouTube)**
-  **TikTok**
(direct messages, comments on videos/photos/livestreams)
-  **Twitch** (public chat, private messages)



Twitter is not a service used by large numbers of young people, however we recognise that we need policies to protect against this... We continue to monitor the development of technology... but currently it is not of sufficient capability or accuracy to be deployed on Twitter.'

Twitter response to the notice question about measures in place to prevent grooming on Twitter.



11% of young people aged 14-17 report they had been asked on the internet for a photo or video showing their private parts when they didn't want to.'

eSafety Commissioner 2022

[Mind the Gap | eSafety Commissioner](#)



In 2022, we saw a 60% increase in the number of images including children aged 7-10 years old. As ever-younger children become more tech-aware and active online, they become more vulnerable to grooming and abuse by strangers – even in their own bedrooms.'

Internet Watch Foundation (IWF)

[IWF Annual Report 2022](#)

Blocking URLs linking to known child sexual exploitation and abuse material

eSafety investigators are aware of platforms being used to distribute thousands of links to child sexual exploitation and abuse sites.

Despite the availability of databases that identify URLs which link to known child sexual exploitation and abuse material and websites that are dedicated to it, some companies are not using them.



Every URL on the list depicts indecent images of children, advertisements for or links to such content. The list typically contains 5,000 URLs but is subject to fluctuation. The list is updated twice a day to ensure all entries are live. As well as making the internet a safer place for everyone, this initiative can help to diminish the re-victimisation of children by restricting opportunities to view their sexual abuse and may disrupt the accessibility and supply of images to those who seek them out.*

Internet Watch Foundation (IWF)
The IWF provides a list of website URLs which they have confirmed contain images and videos of child sexual abuse. [URL list policy](#)

Use of databases to identify URLs linking to known child sexual exploitation and abuse

Not using:



Discord
(public and private servers, direct messages)



Google*
(YouTube, Drive, Meet, Chat, Photos, Messages, Gmail, Blogger)



TikTok
(direct messages**)

Using:



TikTok
(on photos/videos, in photo/video descriptions)



Twitch



Twitter
(Tweets, direct messages)

* If Google detects links to known CSEA it de-indexes them from surfacing on Google search. Google also reports them to NCMEC and hash-matches any CSEA content.

** TikTok stated that it is planning to roll out on direct messages in 2023.

Community moderation

In user-governed online communities, some service providers use appointed volunteers to actively support content moderation and enforcement of community rules, alongside the tools and resources deployed by the service itself.

These volunteer roles, such as creators, streamers, administrators and moderators, are given administrative power to remove unacceptable material and ban violators. Where there is no standards enforcement policy that outlines the responsibilities and expectations of these volunteers, enforcement of rules can be inconsistent, including with regard to child sexual exploitation and abuse. Some self-appointed creators, streamers, and administrators or moderators can also set up dedicated channels for the exploitation and abuse of children.

In addition, a lack of engagement between volunteer moderators and the Trust and Safety staff of a service increases the risk of sexual predators continuing to abuse and re-victimise children, because they may only be banned from a specific channel or group, rather than the whole service.

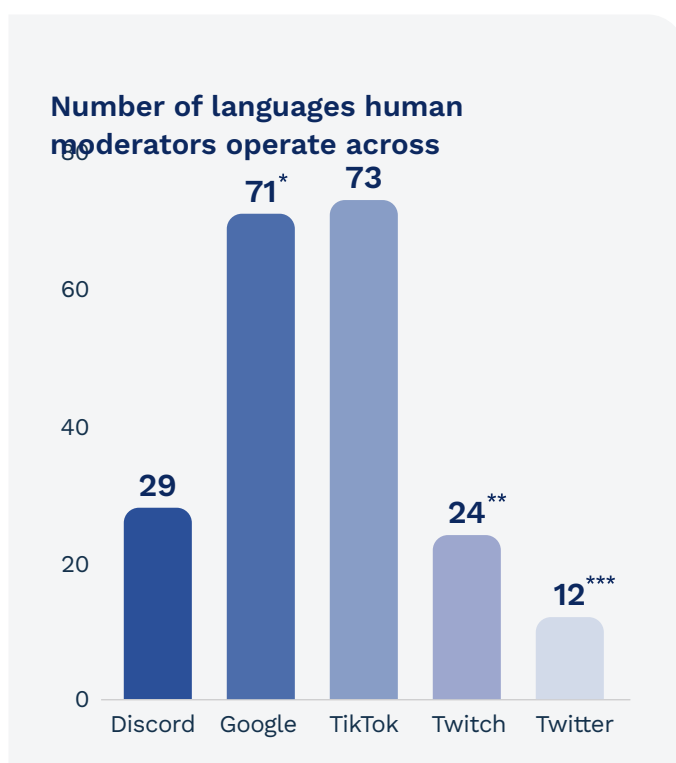


- Discord does not have a standards enforcement policy outlining the responsibilities and expectations of volunteer moderators or administrators
- Twitch does have a standards enforcement policy.
- Professional Trust and Safety staff at Discord and Twitch are not automatically notified when volunteer moderators, administrators or creators take action against child sexual exploitation and abuse material.
- Users on Discord and Twitch are not able to directly report volunteer moderators, administrators or creators for failing to enforce rules.

Language coverage

Review by human moderators may be required to verify abusive content flagged by technology or reported by users particularly where the material has not been identified previously. This is true for grooming, as well as other harmful content such as hate speech, where effective moderation depends on understanding the language and culture for context.

It's important that service providers have human moderators operating in the languages of the communities that use their services.



* Google stated that it operates across at least 70 languages and its 'language capability may vary at any time.'

** Twitch noted that additional languages can and will be implemented when there is a need through additional third-party vendors.

*** Twitter stated that 'The company has the ability to seek and conduct vendor services in a range of additional languages, which include but are not limited to those needed in the event of additional reviews, or for emergencies.'

User reporting

When illegal content such as child sexual exploitation and abuse material is reported by a user, verifying it and taking action should be done quickly to prevent ongoing or new harm.

The responses to the notices highlighted significant differences in the time service providers take to consider and respond to user reports about child sexual exploitation and abuse material.

Provider	Service or parts of service	Median time for user reported CSEA material to be actioned by the provider
Google	Drive, Meet, Chat, Google Photos, Gmail, Blogger, YouTube, Google Messages	Google did not provide the median time from the point when a user makes a report to the report being actioned.
Twitter	Public posts ('tweets') Direct messages	Twitter did not respond.
TikTok	Photos/videos shared publicly (to everyone or followers only)	5.2 minutes
	TikTok Live	7.7 minutes
	Direct Messages	7.4 hours
Twitch	Twitch	8.22 minutes
Discord	Direct messages	13 hours
	Servers (public)	8 hours
	Servers (private)	6 hours
	Server Livestreams	Discord stated it was unable to calculate the response time as there is no in-service reporting option.



Survivors surveyed by C3P have generally characterized their own experience reporting CSAM [child sexual abuse material] online as disheartening; exceedingly long delays in responding to their complaints, moderators challenging victims on the veracity of their report or, as is often the case, no response at all.'

The Canadian Centre for Child Protection (C3P) 2020
protectchildren.ca

Scale of abuse

Although the full size and scope of online child sexual exploitation and abuse is difficult to measure, eSafety knows from the experience of our own investigators and other expert organisations that the scale is significant.



In 2022, ESPs submitted 49.4 million images to the CyberTipline of which 18.8 million (38%) were unique. Of the 37.7 million videos reported by ESPs, 8.3 million (22%) were unique.

National Centre for Missing and Exploited Children (NCMEC) CyberTipline Data 2023
missingkids.org



In “2022...the IWF and partners blocked 8.8 million attempts in one month to access known child sexual abuse material”

Internet Watch Foundation (IWF)
[IWF Annual Report 2022](#)

Note: NCMEC’s CyberTipline is the USA’s centralised reporting system for the online exploitation of children. In 2022, the CyberTipline received 31.8 million reports from Electronic Service Providers (ESPs) about apparent child sexual abuse material on their services.

Child sexual exploitation and abuse reports sent to NCMEC's Cybertipline 2022 (for providers that received notices)



It is important to note that a high reporting figure may indicate the provider is taking the issue seriously and has technology and processes in place to detect and report online child sexual exploitation and abuse material and video livestreaming of abuse on its platforms and services. eSafety is therefore concerned by providers reporting low numbers despite their platforms and services being widely and commonly used.

We can only know the true scale of the global problem if all online services use readily available technologies and human moderation to detect child sexual exploitation and abuse material, video livestreaming of abuse, grooming of children and sexual extortion.

As well as their regulatory requirements in Australia - Twitter, Google, TikTok and Amazon (parent company of Twitch) have endorsed the [Voluntary Principles to Counter Online Child Sexual Exploitation and Abuse](#) and in doing so have committed to taking steps to detect known and new CSEA, including grooming and livestreaming as part of the implementation of those principles. The principles also include steps to share 'meaningful data and insights' on their implementation.

Information obtained in response to eSafety's non-periodic notices can therefore help eSafety understand how providers are implementing the Voluntary Principles.

[^]Discord, 'An update on our business', 2021, accessed 19 January 2023, URL: <https://discord.com/blog/an-update-on-our-business>

^{^^}TikTok, 'Thanks a billion!', September 2021, accessed 17 February 2023, URL: <https://newsroom.tiktok.com/en-us/1-billion-people-on-tiktok>

^{*}Twitch, 'Press centre', 2023, accessed 10 February 2023, URL: <https://www.twitch.tv/p/press-center/>

^{**}[Twitter Inc Annual Report 2021](#) (annualreports.com)

No average user volumes available for all Google services

4. 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 safer online. The current determination was registered on 23 January 2022.

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

Further information on the Expectations and associated powers can be found in eSafety's [Regulatory Guidance](#).

5. Information about the Notices

5.1 Who received Notices?

In February 2023, eSafety issued Notices to five providers under section 56(2) of the Act. The Notices covered the period 24 January 2022 to 31 January 2023 and required each provider to respond to specific questions and report on its implementation of specific Expectations listed in the Notices as they relate to CSEA, and in some cases sexual extortion and the safety of recommender systems.

The providers and services captured by the Notices are set out below.

Provider that received the section 56(2) Notice	Services
Google LLC.	Google Drive (Drive) Google Photos (Photos) Google Meet (Meet) Google Chat (Chat) Gmail Google Messages (Messages) Blogger YouTube

Provider that received the section 56(2) Notice	Services
Twitter Inc. (subsequently X Corp.)	Twitter
TikTok Pte. Ltd.	TikTok
Twitch Interactive Inc.	Twitch
Discord Inc.	Discord

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

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

Examples of other matters that eSafety has said in the 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 material accessible in Australia which are more, or less, compliant with the Expectations than the providers who received Notices.

5.2 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.

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.

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

Areas covered by Notices	Corresponding Expectations in the Determination
The extent to which 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) Section 11 (Minimising Provision of Certain Material)

Areas covered by Notices	Corresponding Expectations in the Determination
Steps taken to prevent and detect livestreamed child sexual exploitation and abuse	Section 6 (Ensuring Safe Use) Section 11 (Minimising Provision of Certain Material)
Steps taken to prevent and detect the grooming and sexual extortion 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) Section 8 (Encrypted Services) 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) Section 9 (Anonymous Services) 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) Section 9 (Anonymous Services) Section 11 (Minimising Provision of Certain Material) Section 14 (Policies and Procedures to deal with complaints)
Steps taken to avoid the risk of amplifying harmful content through recommender systems	Section 6 (Ensuring Safe Use) Section 11 (Minimising Provision of Certain Material)
Availability of mechanisms for users to report child sexual exploitation and abuse material or activity	Section 6 (Ensuring Safe Use) Section 11 (Minimising Provision of Certain Material) Section 13 (Mechanisms to report and make complaints about certain material) Section 14 (Policies and Procedures to deal with complaints) Section 15 (Mechanisms to report and make complaints about breaches of terms of use)

5.3 What was the Notice process?

Providers had 35 days to respond, or longer as agreed with eSafety. Providers were invited to discuss with eSafety any questions they might have about the notices, how to respond, or about the scope of the questions. Several extensions were granted where requested by providers.

5.4 What process was followed once the information was received?

Assessment and Follow-up Questions

On receipt of responses eSafety assessed if each provider had answered the questions required by the Notice. Where a provider did not provide a response (for example, leaving a response to a question in the Notice blank), did not provide a full response (for example, not answering each part of the question required by the Notice), or did not respond using the template provided in the Notice, eSafety followed up with the provider to:

- Give a further opportunity to provide the information sought in the Notice;
- Understand the reasons why the provider was not able to provide the information in answer to a question in the Notice;
- Help eSafety assess the provider's compliance with the Notice.

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

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

Draft Summary Reports

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

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

As outlined in section 5.4 this report provides a summary of the information that eSafety received from the five 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 and to prevent recidivism. 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 period in time – the period 24 January 2022 to 31 January 2023 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 January 2023, 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 in this report is intended to be read alongside the [report published in December 2022](#) to give an overview of the interventions in place on a wider range of services. However, the different reporting periods mean that information may have changed, and new interventions put in place or withdrawn since that report.

The information presented in this summary provides new insight into the steps that these providers are taking to address online CSEA. eSafety hopes that the information obtained from Notices on the Expectations will be used by all industry participants to address key challenges, and incentivise greater transparency in the future, including through their own voluntary disclosures. eSafety also intends that the information in this report, together with the December 2022 report, is used by researchers, academics, the media and public to scrutinise the efforts of industry, to improve accountability and encourage implementation of the Expectations.

5.6 What happens next?

- eSafety will use the information gathered from these Notices to continue 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's regulatory guidance, first published in July 2022 and updated in September 2023, sets out eSafety's planned approach to the exercise of its powers in respect of the Expectations. In the coming months eSafety will:
 - Expand use of non-periodic notices to other acute harms and welcomes input from all stakeholders on the areas where greater transparency is needed.
 - Issue the first periodic notices to begin tracking key safety issues and progress against them. These are likely to focus on acute harms in the first instance, and potential deficiencies in providers safety processes.
 - Consider whether to issue statements of compliance and non-compliance with the Expectations now that further regulatory guidance has been published.

6. Non-compliance with the Notices and action taken by eSafety

6.1 eSafety's powers to require reports on compliance with the Basic Online Safety Expectations

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

A non-periodic reporting notice requires the provider to prepare a report about compliance with one or more of the Expectations, prepare the report in the manner and form set out in the notice, and to provide it to eSafety⁹. Providers are required to comply with a notice to the extent they are capable.¹⁰

When eSafety issued the Notices in February 2023 eSafety provided a response template as part of each Notice, with specific questions tailored to each provider and service, related to specific Expectations. The Notices required responses to these questions.

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

6.2 Findings of non-compliance

eSafety found that Google and Twitter did not comply with the Notices given to them in February 2023.

Google and Twitter failed to provide the information required in response to a number of questions in the Notices. In doing so, they failed to prepare a report in the manner and form required to the extent they were capable and contravened section 57 of the Act.

⁹ Section 49(2)(b) and (c) and Section 56(2)(b) and (c) of the Act

¹⁰ Section 50 and Section 57 of the Act

Google

eSafety has found that Google did not comply with its Notice for the following reasons:

- Google provided answers in certain instances which were not relevant, or generic, where specific information was sought.
- Google provided aggregated information across multiple services, where information regarding specific services was required.

Google submitted that it was not capable of providing certain information in the time available but did not engage with eSafety during the notice period to seek additional time or clarification that might have enabled compliance.

eSafety advised Google that it had failed to answer certain questions and gave further opportunity to provide the information or reasons that the information could not be provided. In several instances Google subsequently provided the information required by the Notice.

eSafety has found that Google did not comply with the Notice to the extent it was capable.

Google has been given a formal warning, notifying it of its failure to comply, and warning it against non-compliance in the future.

Twitter

eSafety found that Twitter did not comply with its Notice for several reasons:

- In some instances Twitter failed to provide any response to the question, such as by leaving the boxes entirely blank.
- In other instances, Twitter provided a response that was otherwise incomplete and/or inaccurate.¹¹

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

eSafety advised Twitter that it had failed to answer certain questions and gave further opportunity to provide the information or reasons the information could not be provided. In several instances Twitter subsequently provided the information required by the Notice.

¹¹ Following publication of this Transparency Report X Corp. advised eSafety that its response to questions regarding the detection of livestreaming of CSEA on the Twitter service 'was inaccurate due to an inadvertent error' and provided a revised response to the question. X Corp.'s revised response is [available](#)

eSafety has found that Twitter did not comply with the Notice to the extent it was capable, and considers its failure to comply to be more serious than Google's for the reasons set out above.

eSafety has given a service provider notification to Twitter, confirming its non-compliance, as well as an infringement notice for \$610,500. Twitter has 28 days to request the withdrawal of the infringement notice or to pay the penalty. If Twitter chooses not to pay the infringement notice, it is open to the Commissioner to take other action.

6.3 Why it is important that providers comply with reporting Notices

Providers are required to comply with their legal obligations under Australian law, including the Act.

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

The Notices related to the most serious and acute harm – the sexual exploitation and abuse of children.

It is providers themselves who hold the information about the internal systems, processes and tools they use to detect and address child sexual exploitation and abuse, and other online harms. The Act recognises this and provides eSafety with powers to mandate the provision of information. The enforcement powers, including civil penalty powers, are considerable, reflecting the importance of compliance.

eSafety spent considerable time and resource subsequently seeking information from providers that should have been provided as required by the Notices. This has obstructed, limited and delayed public transparency and accountability, and prevented eSafety from delivering its functions under the Act.

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

7. Transparency: Responses by Issue

Where providers reported on information that addressed the same or similar issues, eSafety has compiled that information in summary tables commencing at page 30. 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 operate. This information can also be reviewed in light of similar questions in the [December 2022 report](#). Both reports give a good indication of the range of measures on these services since the commencement of the Act.

eSafety recognises that each provider is different – with different functionality, architectures, business models and user bases. This means an intervention or tool which may be proportionate and appropriate on one service, may not be on another. Accordingly, when reviewing the tables, the nature of the service and the context in which the service operates, and the risk of online harms associated with that service 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 intends 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.

7.1 Detecting previously confirmed ('known') images

Providers were asked about their use of tools to detect images containing known CSEA material. '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 billion¹². 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.

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.

Table 1: Use of hash matching tools to identify known CSEA images

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

Provider	Services	Uses image hash matching tools on these services	
		Yes / No	Names of tools used
Google	Drive (stored content) Drive (shared content) Chat Google Photos Gmail Blogger	Yes (Consumer versions of Drive, Chat and Gmail)	PhotoDNA SHA256 Other proprietary technologies
	Meet Google Messages	No	N/A
Twitter	Public posts ('tweets') Direct messages	Yes	PhotoDNA Safer by Thorn ¹³ Internal tools including, ProfileViewer2 as well as internal machine learning

¹² Farid H (2019) '[Statement to the House Committee on Energy and Commerce Hearing: Fostering a Healthier Internet to Protect Consumers](#)', accessed 11 September 2023

¹³ Following a subsequent question from eSafety, providing Twitter with a further opportunity to provide the information sought in the Notice, Twitter provided the name of this tool.

Provider	Services	Uses image hash matching tools on these services	
		Yes / No	Names of tools used
			models and classifiers that do not have names ¹⁴
TikTok	Profile pictures Public photos Photos in direct messages ¹⁵	Yes	PhotoDNA Internal TikTok model ¹⁶
	Private/Only me photos	No ¹⁷	N/A
Discord	Direct messages Servers (public & private)	Yes	PhotoDNA Contrastive Language-Image Pretraining (CLIP) ¹⁸
	Profile pictures ('avatars')	No	N/A

7.2 Detecting previously confirmed ('known') videos

Providers were asked whether they were deploying hash matching tools to detect videos containing known CSEA material. PhotoDNA for Video, Google's CSAI Match and Facebook's TMK+PDQF are examples of existing tools, made available to organisations. As video material is increasingly ubiquitous online, these tools have become more important than ever to assist in preventing the dissemination of CSEA material. In March 2022, the Tech Coalition, a CSEA industry organisation, announced work to improve the standardisation of video hashing tools used across industry¹⁹.

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.

¹⁴ Following a subsequent question from eSafety, providing Twitter with a further opportunity to provide the information sought in the Notice, Twitter provided this information. eSafety notes these are not hash matching tools.

¹⁵ Relating to the sharing via direct messages of images that have already been posted to the TikTok platform and the settings of the post are not set to private. TikTok clarified that users are not able to upload images (or videos) to direct messages directly from their camera roll.

¹⁶ Consisting of multiple TikTok machine learning algorithms that review videos, images, audio and text-based content and TikTok stated was conceptually equivalent to hash matching.

¹⁷ TikTok stated that it was in the process of implementing hash matching on private/only me content. It stated that all content (including photos posted in private) that is uploaded to the platform is reviewed by 'automated detection technology' (including for CSEA material). Following this initial response, TikTok subsequently stated, 'TikTok has implemented hash matching tools over photos posted in the 'Private'/'Only Me' setting and is currently troubleshooting its use in this part of the service.'

¹⁸ AI model developed by OpenAI. eSafety notes this is not a hash matching tool.

¹⁹ Tech Coalition 2022, [Tech Coalition | Initial Results of the Tech Coalition Video Hash Interoperability Alpha Project \(technologycoalition.org\)](https://techcoalition.org/), accessed 11 September 2023.

Table 2: Use of hash matching tools to identify known CSEA videos

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
Google	YouTube Drive (stored content) Drive (shared content) Google Photos Blogger	Yes (Consumer version of Drive)	CSAI Match
	Meet Chat Google Messages Gmail	No	N/A
Twitter	Public posts “tweets” Direct messages	No	N/A
TikTok	Public videos Videos in direct messages ²⁰	Yes	PhotoDNA Google content Safety API ²¹ CSAI Match ²² Internal TikTok Model ²³
	Private/only me videos	No ²⁴	N/A
Twitch	Twitch	No	N/A
Discord	Direct messages Servers public & private	No	N/A

²⁰ Relating to the sharing via direct messages of videos that have already been posted to the TikTok platform.

²¹ A free toolkit developed by Google to help increase the capacity to review CSEA content in a way that requires fewer people to be exposed to it.

²² TikTok Footnote: This API helps identify re-uploads of previously identified CSAM in videos, using matches against Youtube's database.

²³ TikTok reported that this includes machine learning algorithms that review videos, images, audio and text-based content and stated it was conceptually equivalent to hash matching.

²⁴ TikTok stated that it was in the process of implementing matching on private/only me content.

7.3 Blocking links (URLs) to known CSEA material

eSafety investigators are aware of platforms being used to distribute thousands of links to CSEA sites. Some of these links are posted under the guise of legal adult porn, but they are also explicitly posted and advertised as CSEA.

Providers were asked if they block URLs²⁵ linking to known CSEA material across all parts of their service.

Not-for-profit organisations such as the Internet Watch Foundation (IWF) have databases of known URLs, colloquially known as web addresses or ‘links’, of previously reported images and videos of CSEA, that they provide to Industry. Law enforcement agencies around the world also have databases of URLs to known abuse material. The blocking of URLs is also a common practice across many online services for safety, security or legal reasons.

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

Table 3: Blocking and source of URLs linking to known CSEA material

The following information was given by service providers regarding the blocking of URLs linking to known CSEA material on their service(s) and from where the providers source the URLs.

Provider	Services	Blocks URLs to known abuse material	
		Yes / No	URL databases/sources used
Google	YouTube	No	N/A
	Drive (stored and shared content)		
	Meet		
	Chat		
	Google Photos		
	Google Messages		
	Gmail		
	Blogger		

²⁵ Uniform Resource Locator. An example of a URL list of known CSEA is the [Internet Watch Foundation's URL List](#), accessed 11 September 2023

Provider	Services	Blocks URLs to known abuse material	
		Yes / No	URL databases/sources used
Twitter	Public posts “tweets” Direct messages	Yes	<ul style="list-style-type: none"> • NCMEC • IWF • Law enforcement • Cybertip.ca²⁶ • Agents internal CSE investigations
TikTok	Links overlayed in photos/videos Links in photo/video descriptions	Yes	<ul style="list-style-type: none"> • IWF • Internal TikTok Library²⁷
	Links sent in direct messages	No ²⁸	N/A
Twitch	Twitch	Yes	<ul style="list-style-type: none"> • Internal investigations and language analysis • Crisp²⁹ • Cross-industry sharing of URLs³⁰
Discord	Direct messages Servers public & private	No	N/A

7.4 Detecting new material

Hash matching tools can only ‘match’ against previously identified and confirmed (‘known’) CSEA material and seek to prevent its ongoing dissemination. But steps can also be taken to prevent the sharing of CSEA material when it is first created, and before it has been

²⁶ Following a subsequent question from eSafety, providing Twitter with a further opportunity to provide the information sought in the Notice, Twitter provided this additional information.

²⁷ Captured from prior violations.

²⁸ TikTok stated that it will be launching a product feature in September 2023.

²⁹ Following this initial response, Twitch subsequently provided eSafety with further information regarding URL sources, and stated, ‘Twitch is constantly investing in, measuring, and optimizing its overall safety and content moderations programs. As a result, the precise constituent tools of those programs and our reliance on those tools is frequently in flux. For example, while Twitch relied in several areas on the third party Crisp at the time of its submission, Twitch has since significantly scaled down its reliance.’

³⁰ Through ‘Threatexchange’, Tech Coalition, Thorn, and others.

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 material, in order to identify the markers of content depicting abuse. An example of this technology is Google’s Content Safety API³¹ or Thorn’s classifier, which Thorn reports has a 99% precision rate.³²

In 2022, INHOPE, an international network of CSEA hotlines, reported that ‘84% of CSAM reports were of never-before-seen material, depicting new and known child victims of sexual abuse.’³³

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

Table 4: Use of tools to identify new CSEA material

The following information was given by service providers regarding tools being used to detect new CSEA material.

Provider	Services	Uses tools on these services to identify new CSEA material	
		Yes / No	Names of tools used
Google	YouTube Drive (stored content) Drive (shared content) Google Photos Blogger	Yes (Consumer version of Drive)	• Machine learning classifiers (known externally as Google Content Safety API)
	Meet Chat Google Messages Gmail	No	N/A

³¹ Google, ‘[Child safety toolkit](#)’, accessed 11 September 2023.

³² Thorn (2020), ‘[How Safer’s detection technology stops the spread of CSAM](#)’, accessed 11 September 2023

³³ INHOPE Annual Report 2022, [INHOPE | INHOPE Annual Report 2022](#), accessed 11 September 2023

		Uses tools on these services to identify new CSEA material	
Provider	Services	Yes / No	Names of tools used
Twitter	Public posts “tweets”	Yes	<ul style="list-style-type: none"> • Safer by Thorn³⁴ • Several internal tools including, ProfileViewer2 and internal machine learning models and classifiers that do not have names.³⁵
	Direct Messages	No	N/A
TikTok	Profile pictures Public photos Private/only me photos Images/Video in direct messages ³⁶ TikTok live video	Yes	<ul style="list-style-type: none"> • Internal computer vision model³⁷ • Audio and NLP (Natural Language Processing) model • Google Content Safety API
Twitch	Twitch	Yes	<ul style="list-style-type: none"> • Stream scanner – internally developed tool • Crisp’s proprietary open-source AI tech³⁸ • Groomer detection tools - Automated heuristics to identify and report grooming
Discord	Direct messages Servers (public & private)	Yes	<ul style="list-style-type: none"> • Contrastive Language-Image Pretraining (CLIP)
	Profile pictures (avatars)	No	N/A

³⁴ Following a subsequent question from eSafety, providing Twitter with a further opportunity to provide the information sought in the Notice, Twitter provided the name of this tool.

³⁵ Following a subsequent question from eSafety, providing Twitter with a further opportunity to provide the information sought in the Notice, Twitter provided this information.

³⁶ Relating to the sharing via direct messages of images or videos that have already been posted to the TikTok platform and the image or video was not posted in the ‘Private’/‘Only Me’ setting. TikTok clarified users are not able to upload images or videos directly from their camera roll.

³⁷ TikTok defined this as a model that divides videos into frames, which are then processed (for example, to detect exposed body skin).

³⁸ See previous footnote regarding Crisp.

7.5 Proactive detection

eSafety asked what portion of CSEA material found on the various services was detected proactively by various technologies compared with material reported by a user or trusted flaggers. Proactive detection rates can provide an indication of the extent that technology is used on a service to detect CSEA.

A service that does not deploy such tools, or uses proactive detection technology, systems and processes less extensively, will likely detect less CSEA. Such a service will typically detect CSEA through a higher proportion of user reports.

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

Table 5: Proactive detection of CSEA material

The following information was given by service providers regarding proactive detection of CSEA on their services versus user or trusted flagger detection of CSEA.

Provider	Services	Proportion of proactive detection of CSEA (versus user/trusted flagger reporting)
Google	YouTube Drive Chat (consumer version) Google Photos Gmail (consumer version)	99%
	Blogger	>99%
	Meet Google Messages	No automated detection tools used on these services so Google stated it could not provide the same comparison.
Twitter	Public posts “tweets” Direct Messages	24 January 2022 – 27 October 2022 ³⁹ 90%
		28 October 2022 – 31 January 2023 75%

³⁹ eSafety sought answers on two separate time periods to reflect Twitter’s acquisition.

Provider	Services	Proportion of proactive detection of CSEA (versus user/trusted flagger reporting)
TikTok	TikTok	94.4% ⁴⁰
Twitch	Twitch	~20% ⁴¹
Discord	Direct messages	72.1%
	Servers (public)	87.8%
	Servers (private)	44.5%

7.6 Detecting CSEA in livestreams or video calls

The Australian Institute for Criminology (AIC) has published analysis of offender chat logs showing the use of 'popular platforms to arrange and watch the sexual abuse of children live'.⁴² A further report by the AIC suggests that 'some CSA live streaming offenders generate their own CSAM through live streaming sessions' and that 'It was common for offenders...to request images and/or videos of the victims they viewed in the CSA live streaming sessions'.⁴³ This 2023 Report by the AIC concludes by stating that the, 'scanning of live streams for abusive content, should be a priority for all electronic service providers.'

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.⁴⁴ 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.⁴⁵ Sometimes this is in exchange for payment. The International

⁴⁰ Global proactive detection rate of CSEA, excluding the USA

⁴¹ Following its initial response, Twitch subsequently provided eSafety with further information stating that in Q1 2023, ~20% of CSEA material was identified via proactive detection systems and proactive investigation by Twitch's safety team. Twitch clarified that this figure does not include the prevention of material occurring on the platform in the first place.

⁴² Australian Institute of Criminology (2021), "[Live streaming of child sexual abuse: An analysis of offender chat logs](#)", accessed 11 September 2023.

⁴³ Australian Institute of Criminology (2023), [The overlap between child sexual abuse live streaming, contact abuse and other forms of child exploitation \(aic.gov.au\)](#), accessed 11 September 2023.

⁴⁴ IICSA (2019) '[IICSA Inquiry - Internet Hearing](#)', accessed 11 September 2023.

⁴⁵ Australian Federal Police (2021) [AFP warn about fast growing online child abuse trend | Australian Federal Police](#) accessed 11 September 2023.

Justice Mission describes livestreamed CSEA material as ‘live crime scenes happening on tech platforms.’⁴⁶

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 service provider Yubo has published information⁴⁷ 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 published⁴⁸ 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 prompts⁴⁹ or age assurance measures.

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

⁴⁶ 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 11 September 2023.

⁴⁷ Yubo, (n.d.) ‘[Safety Hub](#)’, accessed 11 September 2023.

⁴⁸ Camera Forensics (2022) ‘[Understanding real-time threat detection and the SafeToWatch mission](#)’, accessed 11 September 2023.

⁴⁹ The Lucy Faithful Foundation, Sept 2022, [Stop It Now, Internet Watch Foundation and Pornhub launch first of its kind chatbot to prevent child sexual abuse \(lucyfaithfull.org.uk\)](#), accessed 11 September 2023.

Table 6: Detecting CSEA in livestreams or video calls

The following information was given by service providers regarding detecting CSEA material in livestreaming, or video calls/video conferences.

Provider	Services	Measures in place to detect CSEA in livestreams or video calls/conferences – Yes/No	Tools used
Google	YouTube	Yes	<ul style="list-style-type: none"> • machine learning language analysis • video classifiers
Twitter	Twitter	Twitter did not provide an answer regarding whether tools were used to detect livestreaming during the reporting period ^{50 51}	
TikTok	TikTok live video	Yes	<ul style="list-style-type: none"> • internal computer vision model • internal audio and NLP model • machine learning behavioural model
Twitch	Twitch	Yes	<ul style="list-style-type: none"> • internal language analysis tools • Spirit AI • Stream scanner (AWS Rekognition) • Crisp⁵²
Discord	Discord	No	N/A

7.7 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. In 2022, 78% of the websites the Internet Watch Foundation (IWF) removed contained images

⁵⁰ Following a subsequent question from eSafety, providing Twitter with a further opportunity to provide the information sought in the Notice, Twitter noted that livestreaming had been deprecated during the reporting period, other than for 'pre-vetted' media partners, but did not provide an answer regarding whether tools were used to detect livestreamed CSEA during the reporting period. It noted that users could report livestreams.

⁵¹ Following publication of this Transparency Report X Corp. advised eSafety that its response to questions regarding the detection of livestreaming of CSEA on the Twitter service 'was inaccurate due to an inadvertent error' and provided a revised response to the question. X Corp.'s revised response is [available](#).

⁵² See previous footnote re Crisp.

or videos where children, some as young as 7, had been groomed and coerced into sexual activities over an internet enabled device with a camera by an abuser⁵³.

Detecting grooming on online services presents some 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 alternative 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.⁵⁴ 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.

For this reason, eSafety also asked services to provide the proportion of reports that their language analysis tools send to a human for moderation. A high percentage sent for human moderation may indicate the service conducts analysis and investigation into likely grooming activity to ensure accuracy, and to also identify linked activity.

A recent industry 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.⁵⁵

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

⁵³ The Internet Watch Foundation, Annual Report 2022, [IWF 2022 CSAM Annual Report | Internet Watch Foundation](#), accessed 11 September 2023.

⁵⁴ NBC News, Jan 2020, '[Microsoft launches tool to identify child sexual predators in online chat rooms \(nbcnews.com\)](#)', accessed 11 September 2023.

⁵⁵ WeProtect (2021) Findings from WeProtect [Global Alliance/Tech Coalition survey of technology companies](#), accessed 11 September 2023.

Table 7: Use of language analysis technology to detect likely grooming activity and proportion of reports sent for human moderation

The following information was given by service providers regarding the language analysis technology used to detect likely grooming activities and the proportion of reports flagged by these tools that are sent for human moderation.

Provider	Services	Uses language analysis technology on these services	Names of tools used	Proportion of reports flagged by tools sent for human moderation
		Yes / No		
Google	YouTube	Yes	Machine learning classifiers	Google stated that this information is not available
	Meet Chat Google Messages Gmail	No	N/A	N/A
Twitter	Public posts “tweets” Direct messages	No	N/A	N/A
TikTok	Comments on photos/videos Comments on TikTok live videos	Yes	Internal TikTok NLP model Language analysis model	<1% ⁵⁶
	Direct Messages	Yes	Internal TikTok NLP model Language analysis model	0% ⁵⁷
Twitch	Public chat	Yes	Automod Crisp ⁵⁸	100%

⁵⁶ TikTok subsequently added that almost 100% of comments that are flagged by technology as potentially violative of policies including online grooming and likely CSEA activity are automatically actioned by technology and consequently, less than 1% are reviewed by human moderators.

⁵⁷ TikTok subsequently added that to reach a balance between privacy obligations and user safety, whilst reports flagged by technology are not sent for human review, other methods are deployed by technology to detect grooming activities or likely CSEA activity. TikTok also stated that 100% of reports made by a user with regard to a direct message are reviewed by human moderators.

⁵⁸ See previous footnote regarding Crisp.

Provider	Services	Uses language analysis technology on these services	Names of tools used	Proportion of reports flagged by tools sent for human moderation
		Yes / No		
			Internal language analysis tools	
	Private messages (whispers)	Yes	Internal language analysis tools	100%
Discord	Direct messages Servers (public) Servers (private)	No	N/A	N/A

7.8 Detecting sexual extortion and other exploitation of children

Sexual extortion, also known as sextortion, is a crime involving online blackmail, where victims are tricked into sending intimate images of themselves to someone who then threatens to share the images unless demands are met, usually for payment. This growing crime is targeting teenage males in particular.

eSafety has noted an increase of sexual extortion reports⁵⁹ from almost 600 reports in the first quarter of 2022 to more than 1,700 in the first quarter of 2023. The majority of these reports were from young people, with 90% of all reports coming from males. Many of these young people report paying thousands of dollars and many more suffer deep distress.

In the period April to December 2022, the Australian Centre to Counter Child Exploitation (ACCCE) saw a 6-fold increase in the volume of sexual extortion cases involving children under 18. Almost 60% of children targeted were 16-17 years old, with 37% of children targeted 13-15 years old. Sexual extortion constitutes CSEA where the victim is under the age of 18.

In November 2022, the Canadian Centre for Child Protection (C3P)⁶⁰ published a [report](#) on the analysis of financial sextortion victim posts. The report explains how extortioners weaponise social media platforms using similar strategies to entice, blackmail and threaten

⁵⁹ eSafety Commissioner, [Media Release, Sexual extortion and child abuse reports almost triple | eSafety Commissioner](#), May 2023, accessed 11 September 2023

⁶⁰ C3P is a national charity dedicated to the personal safety of all children. C3P operates Cybertip.ca, Canada's tipline to report child sexual abuse and exploitation online, [Resources & Research: An Analysis of Financial Sextortion Victim Posts Published on r/Sextortion – protectchildren.ca](#), accessed 11 September 2023

victims. The report voices victims' concerns about 'platform reporting functions that failed to provide them with options to accurately describe their situation and a lack of meaningful action being taken by platform operators.' The report also lists functionalities that make it easy for offenders to connect to victims, functionalities that incentivise users to share personal information, and the apparent ease of offenders to create multiple fraudulent accounts as additional platform design risks contributing to the proliferation of this form of child sexual exploitation.

In addition to the sexual extortion of children, reports from media outlets Protocol⁶¹, Forbes⁶², Financial Times⁶³ and NBC News⁶⁴ indicate that offenders use key codewords and signals with thinly veiled references to CSEA to advertise, exchange, trade, sell or distribute child sexual exploitation material. In a report looking at the cross-platform dynamics of self-generated CSAM (SG-CSAM)⁶⁵, including sextortion, the Stanford Internet Observatory (SIO) expressed the view that platforms should have 'better proactive investigations and heavier enforcement on keywords and hashtags' and should share information between platforms and services on 'changing characteristics, hashtags, [and] keywords' amongst other recommendations, to improve detection of this type of child exploitation at an industry-wide level.⁶⁶

Techniques are available to industry to detect and address the issue. One example is the availability of the UK's Internet Watch Foundation's (IWF) 'keyword list', compiled of words, phrases and codes associated with various forms of child sexual exploitation. This list is updated monthly. Language tools using such lists, and lists developed by industry, can detect the key words, terms, abbreviations, codes and hashtags associated with the above forms of child sexual exploitation.

eSafety asked about the detection of likely terms, abbreviations, codes and hashtags indicating likely CSEA activity, in particular but not limited to sexual extortion and the sale and trade of CSEA, in relation to section 6(2) of the Determination, and section 11.

⁶¹ Protocol, [TikTok is under investigation over spread of child sexual abuse material](#), April 2022, accessed 11 September 2023

⁶² Forbes, [These TikTok accounts are hiding child sexual abuse material in plain sight](#), November 2022, accessed 31 May 2023

⁶³ Financial Times, [TikTok under US government investigation on child sexual abuse material](#), April 2022, accessed 11 September 2023

⁶⁴ NBC News, [On Musk's Twitter, users looking to sell and trade child sex abuse material are still easily found](#), January 2023, accessed 11 September 2023,

⁶⁵ Stanford Internet Observatory (SIO) define SG-CSAM as child sexual abuse material that has been self-generated, that is, 'an image or video [that] appears to be created by the minor subject in the image.' SIO also say SG-CSAM 'can overlap with non-consensual intimate imagery', where 'sexual images are distributed without the consent of the subject', and 'sextortion'. Stanford Internet Observatory, Cyber Policy Centre, [Cross-platform Dynamics of Self-Generated CSAM](#), June 2023, accessed 11 September 2023.

⁶⁶ Stanford Internet Observatory, Cyber Policy Centre, [Cross-platform Dynamics of Self-Generated CSAM](#), June 2023, accessed 11 September 2023.

Table 8: Use of language analysis technology to detect likely CSEA activity such as sexual extortion and other exploitation of children

The following information was given by service providers regarding the language analysis technology used to detect likely CSEA activity, where terms, codes, abbreviations and hashtags are sourced, and the proportion of reports flagged by these tools that are sent for human moderation.

Provider	Services	Uses language analysis technology to detect terms/ codes, etc on these services	Organisations where terms/codes, etc are sourced	Proportion of reports flagged by tools sent for human moderation
		Yes / No		
Google	YouTube	Yes	Does not source terms from external organisations	100% (all content flagged as having reached a certain threshold of likelihood of being CSEA)
	Meet Chat Google Messages Gmail	No	N/A	N/A
Twitter	Public posts “tweets”	Yes	Twitter stated that it was unable to disclose the specific organisations it uses to source terms and abbreviations, due to confidentiality agreements with the providers Twitter works with.	‘The majority’ ⁶⁷

⁶⁷ Twitter stated that the majority of accounts flagged for CSEA are reviewed by a human moderator, but that it had a small number of high precision signals where it suspends users before a human reviews it in order to quickly act on violating accounts.

Provider	Services	Uses language analysis technology to detect terms/ codes, etc on these services	Organisations where terms/codes, etc are sourced	Proportion of reports flagged by tools sent for human moderation
		Yes / No		
	Direct Messages	No	N/A	N/A
TikTok	Comments on photos/ videos	Yes	IWF Library NGOs (Thai government agency, IJM Philippines)	<1% ⁶⁸
	Comments on TikTok live videos		TikTok in-house intelligence sources Internal TikTok Library	
	Direct Messages	Yes	As above	0% ⁶⁹
Twitch	Public chat	Yes	Internal teams	100%
	Private messages (whispers)		Amazon teams Other organisations such as Thorn	
Discord	Discord	No	N/A	N/A

7.9 Detecting underage users

Many providers' terms of service state that 13+ is the minimum age to use their services, yet a 2020 eSafety report on the digital lives of Australian teenagers⁷⁰ found that up to 66% of 12- to 13-year-olds were using YouTube and that 12- to 13-year-olds use on average 3.1 social media services. This suggests that it is common for young people to not provide

⁶⁸ TikTok subsequently added that almost 100% of comments that are flagged by technology as potentially violative of policies including online grooming and likely CSEA activity are automatically actioned by technology and consequently, less than 1% are reviewed by human moderators.

⁶⁹ TikTok subsequently added that to reach a balance between privacy obligations and user safety, whilst reports flagged by technology are not sent for human review, other methods are deployed by technology to detect grooming activities or likely CSEA activity. TikTok also stated that 100% of reports made by a user with regard to a direct message are reviewed by human moderators.

⁷⁰ eSafety Commissioner, [The digital lives of Aussie teens](#), February 2021, accessed 11 September 2023.

accurate age information when they sign up and that platforms are not detecting this misuse of their services.

In a 2022 report, Forbes⁷¹ reported on cases of young girls using a livestreaming service and performing sexually suggestive acts in exchange for receiving virtual currency from strangers. Forbes reportedly reviewed hundreds of livestreams and reported observing users urging young girls – including those who self-identified as minors – to perform sexually exploitative acts, and that those who did so were rewarded with gifts or off-platform payments. This suggests that age-restrictions can be circumvented by underage users and this is exploited by offenders to groom children and solicit them into self-generating sexually exploitative material, with the potential for sexual extortion.

There are various methods available to providers to help determine age at the point of sign-up and ways to detect underage users already on the platform, such as:

- Use of behavioural signals to collect data on how a user interacts with a service. This data can help providers make predictions or determinations about a user, including their age. For example, messages can be flagged where users state an age different to the one on their profile.
- Age-gating or self-declaration. If the user indicates that they are under the relevant age, they cannot proceed to the next step of signing up. As noted above, self-declaration relies on the user providing accurate information. Account-based assurance or cross-account authentication allows a user to share attributes such as their age by logging into an existing account, such as their Facebook or Google account, when accessing an age-restricted site.
- Biometric age assessments analyse a person's characteristics, such as their face or voice, to estimate their age.
- Hard identifiers verifies a person's age through an official or authenticated document which is uploaded to a service and the details of this document validated against an existing database. In instances where a photo ID is used, users may also be required to share a real-time photo of themselves so it can be compared with the photo on the ID document.

eSafety asked about the detection of underage users in the context of section 6 (2) and section 11.

⁷¹ Forbes, [How TikTok Live became a 'strip club filled with 150year-olds'](#), April 2022, accessed 11 September 2023

Table 9: Volumes of underage users, tools used to detect underage users and % of underage users reported by users

The following information was given by service providers about the steps they take to detect underage users on their services.

Provider	Service	Number of underage users detected		Interventions used	% detected proactively (compared with identified by user reports)
		Globally	Australia		
Google	YouTube Drive Meet Chat Google Photos Google Messages Gmail Blogger	Approx. 19 million Google Accounts Did not provide breakdown by service	Google stated the data was not available	<ul style="list-style-type: none"> Classifiers scan for signals on YouTube Machine learning systems Requires new accounts to enter date of birth 	Google stated this information was not available
Twitter	Twitter	703,561	7,618	<ul style="list-style-type: none"> Requires new accounts to enter date of birth User reporting 	Twitter provided a number that it subsequently indicated was unrelated to the information sought
TikTok	TikTok	79,312,821	963,158	<ul style="list-style-type: none"> Requires new accounts to enter date of birth User reports Language analysis 	80.6% (global)

Provider	Service	Number of underage users detected		Interventions used	% detected proactively (compared with identified by user reports)
Twitch	Twitch	120,070	2,590	Requires new accounts to enter date of birth Report from parents Text and traffic analysis Report evaluation and analysis	14.8% ⁷² (global)
Discord	Discord	224,000	4,183	<ul style="list-style-type: none"> Requires new accounts to enter date of birth (and additional interventions when there is a failed attempt to enter a valid birthdate) User reports 	2% ⁷³ Discord did not indicate whether figure was global or Australian.

7.10 Detecting recidivism

eSafety investigators regularly see the same offenders create multiple new accounts, even after they have been banned by a platform. A review of image-based abuse reports to eSafety regarding sexual extortion and child sexual exploitation revealed multiple accounts were created using different images of the same individual. Of these accounts, numerous were created using variations of the same account usernames. These accounts were then

⁷² Twitch noted that this did not include those users who were prevented from re-creating an account after being previously identified as underage.

⁷³ On the basis that Discord stated that 98% were identified through user reports.

used to undertake the online harm, targeting both adults and children. In one such instance, there were 22 accounts created using different images of the same individual, all of which had a variation of one of three different account usernames. Of these, 16 accounts were created using different images of the same individual, with each account also using a variation of the same account username. This indicates the accounts were likely to have been created by one offender, or multiple offenders operating together, to perpetrate the online harm. Each of these accounts were created on the same platform and were reported by eSafety to the relevant company for removal.

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 this information being used to avoid detection. Instead eSafety has sought to demonstrate 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 continues to provide service 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 uses the following terms to give an impression of the extent of the indicators used by services in the table below:

- Minimal: A small number.
- Several: A moderate number.
- Multiple: A significant number.

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

- Section 9 expectation that providers will take reasonable steps to have processes in place to prevent anonymous accounts being used to deal with material, or for activity, that is unlawful or harmful.
- 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’.

Table 10: Detecting recidivism

In response to the Notices the following information was given by providers about the steps and indicators taken on their services to prevent recidivism and whether they authenticate user information on sign up.

Provider	Service	Uses steps to prevent recidivism	Number of indicators
Google	YouTube Drive Meet Chat Google Photos Google Messages Gmail Blogger	Yes	Minimal ⁷⁴
Twitter	Twitter	Yes	Multiple
TikTok	TikTok	Yes	Multiple
Twitch	Twitch	Yes	Several
Discord	Discord	Yes	Multiple

eSafety also asked providers whether they authenticate users on sign up to their services in the context of section 6 (2), section 9 and section 14.

Table 11: Authentication of information on sign up

Provider	Service	Authentication of information on sign up Yes / No
Google	YouTube Drive Meet Chat Google Photos Google Messages Gmail Blogger	Yes
Twitter	Twitter	Yes

⁷⁴ Although certain features on YouTube cannot be accessed without additional verification.

Provider	Service	Authentication of information on sign up
		Yes / No
TikTok	TikTok	Yes
Twitch	Twitch	Account verification requirements vary
Discord	Discord	Yes

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. Of the five providers who received Notices in February 2023, only Google has multiple services that eSafety asked about.

Table 12: Cross services ban implemented for CSEA related activity

Provider	Service	Cross-service ban
Google	YouTube	Yes If a user's Google account is terminated for a CSAM violation, they would lose access to other products that are based on or dependent on that Google account.
	Drive (stored content)	
	Drive (shared content)	
	Meet	
	Chat	
	Google Photos	
	Google Messages	
	Gmail	
	Blogger	

7.11 Recommender System safety

Recommender systems, also known as content curation systems, are the systems that prioritise content or make personalised content suggestions to users of online services.

There are many positive outcomes from recommender systems, for example, recommender algorithms that prioritise time spent reading or reacting to a post and then serve up similar content in the future may result in people seeing things they find interesting, entertaining or valuable.

However, There are risks if the objectives of a recommender system is to deliver greater engagement, without regard to safety. Machine learning techniques are often used to identify user attributes and patterns and make recommendations to achieve particular goals, based on a range of data and signals on the service.

A report by the Stanford Internet Observatory (SIO)⁷⁵ found that CSEA, and in particular self-generated CSAM (SG-CSAM), is more efficiently advertised and shared amongst buyers and sellers through recommender systems, ‘Recommendation algorithms inadvertently boost the network; a user who follows one seller account receives related suggestions for others.’

Risks can be greater for children and young people, especially if they are served friend or follower suggestions that encourage them to interact with potentially dangerous adults, content that normalises the sexualisation of young people or content that may be appropriate for adults but harmful to children who are not developmentally ready for it.

A 2019 New York Times report⁷⁶ highlighted the dangers of algorithmically driven recommender systems curating video feeds that promoted videos of children in exposed positions and states of undress. The article reported that many of these videos were inundated with comments from potential offenders communicating with each other and attempting to contact children posting the videos. Commenters were also reportedly observed sharing their contact details to organise CSEA material exchanges offsite.

eSafety asked about recommender systems in the context of sections 6 (2) and section 11 of the Determination.

Table 13: Design objectives of recommender systems

In response to the Notices, the following information was given by providers regarding the design objectives of their recommender systems.

Provider	Service	Design objectives of recommender systems
Google	YouTube	Helping people find the videos they want to watch and that will give them value. Google added that it uses

⁷⁵ Stanford Internet Observatory (SIO), [Cross-Platform Dynamics of Self-Generated CSAM](#), June 2023, accessed 11 September 2023

⁷⁶ The New York Times, [‘On YouTube’s digital playground, an open gate for pedophiles’](#), June 2019, accessed 11 September 2023

Provider	Service	Design objectives of recommender systems
		classifiers and other measures to minimise chances that a viewer will see problematic content.
Twitter	Twitter	<p>To increase time spent on the platform and recommend content with:</p> <ul style="list-style-type: none"> • A high likelihood of engagement (likes, replies, bookmarks) • A low likelihood of being reports of bad content • A low likelihood of author being blocked
TikTok	For You Feed	<ul style="list-style-type: none"> • Providing content that is personalised and of interest to users • Providing diverse content (ie “intersperse diverse types of content along with those users already enjoyed”) • Ensuring safety and quality is maintained for all users
Twitch	Prioritisation of channels in navigation panel	<p>Connecting viewers to streamers and communities that they can engage and be a part of. To achieve this Twitch considers:</p> <ul style="list-style-type: none"> • What other users (that are similar to the user in question) have watched in the past • The viewing patterns of the user in question (based on what channels they interact with and their activity in those channels)
Discord	Server Discovery	<p>Discord said it does not use recommender systems to organise the servers displayed on Server Discovery. Discoverable servers are organised by total number of users.</p>

Table 14: Top recommender system signals

In response to the Notices, the following information was given by providers regarding the top signals their recommender systems use to show content to a user.

Provider	Service	Signals ⁷⁷
Google	YouTube	<p>A number of signals build on each other to help inform the system about what a user finds satisfying:</p> <ul style="list-style-type: none"> • Clicks • Watch time • Survey responses to measure ‘valued watch time’ • Sharing, likes and dislikes <p>Google said the system doesn't follow a set formula, but develops dynamically as a user's viewing habits change and that these signals can be overruled by Google's ‘commitment to meeting [its] responsibility to the YouTube community and to society.’</p>
Twitter	Twitter	<p>Nine relevant components that make up the 'for you' timeline, which have now been published. The types of signals leveraged include:</p> <ul style="list-style-type: none"> • number of likes • number of retweets • views • number of replies • whether a user engaged with the author in the past • users' interests • recency of a tweet • user's sensitive media settings (minors cannot view sensitive no matter their settings) • suspected authenticity ('not spammy, twitter blue')
TikTok	TikTok	<ul style="list-style-type: none"> • User engagement or activity information (such as video playtime, likes, shares, accounts followed, comments, content created) • Account or device information (such as language preference, country setting, device type) • Video information (such as captions, sounds, hashtags) • A strong indicator of interest, such as whether a user finishes watching a longer video from beginning to end, would receive greater weight than a weak indicator, such as whether the video's viewer and creator are both in the same country. <p>TikTok said these weights change dynamically as the models “learn” more about a particular user's behavior. TikTok said to achieve diversification and safety, the recommendation system</p>

⁷⁷ Signals used include, but are not necessarily limited to, those listed.

Provider	Service	Signals ⁷⁷
		incorporates various rules to filter out inappropriate videos or to intersperse content.
Twitch	Twitch	<ul style="list-style-type: none"> Behavioral data such as watch history and history of users with similar behaviors.
Discord	Server Home	<ul style="list-style-type: none"> Overall age of server Overall age of message Number of user reactions to a message Whether a message contains @everyone

7.12 User Reporting

Median time for user reported CSEA material to be actioned

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.

Table 15: Median time for user-reported CSEA material to be actioned

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. material removed, user banned, or other content moderation decision taken).

Provider	Service	Median time for user reported CSEA material to be actioned
Google	Drive Meet Chat Google photos Gmail Blogger YouTube Google Messages	Google did not provide the median time from the point when a user makes a report to the report being actioned. ⁷⁸
Twitter	Public posts “tweets” Direct messages	Twitter did not provide an answer ⁷⁹
TikTok	Photos/videos shared publicly (to everyone or followers only)	5.2 minutes
	TikTok Live	7.7 minutes
	Direct Messages	7.4 hours
Twitch	Users signed into an account	8.22 minutes ⁸⁰
Discord	Direct messages	13 hours
	Servers (public)	8 hours
	Servers (private)	6 hours

⁷⁸ Google instead provided a metric for the time it takes for their child safety team to action reports once they are notified, set out in their individual summary. Google also provided the percentage of content that violates YouTube’s Child Safety policy (which includes but is broader than CSEA) with 10 or fewer views at the time of removal. Following a subsequent question from eSafety, providing Google with a further opportunity to provide the information sought in the Notice, Google noted that providing the ‘time from the point of receiving a user report to the point of responding is complicated’. Google’s reasons are set out in their individual summary.

⁷⁹ Following a subsequent question from eSafety, providing Twitter with a further opportunity to provide the information sought in the Notice, Twitter provided an alternative metric. This is detailed in the individual summary of their response.

⁸⁰ For users signed into the service. Twitch stated that by the time a report of CSEA is reviewed, if the content involved nudity or clearly CSEA video/images (as opposed to text-based CSEA such as grooming), it likely has already been removed by moderation staff. This time reflects enforcement against the account-holder, not actions taken on the content itself.

Provider	Service	Median time for user reported CSEA material to be actioned
	Server Livestreams	Discord stated it was unable to calculate the response time as there is no in-service reporting option.

In-service user reporting

A recent report from Ofcom⁸¹, the UK regulator responsible for online safety, highlighted that where a user does not have the ability to report illegal or harmful activity in-service ‘This creates extra friction and may result in slower, or less reporting.’ Without a readily available, inclusive and easily accessible in-service reporting function users will be less likely to report. Users may not know how to report if there is not an option available alongside the material they are viewing.

In 2020 the Canadian Centre for Child Protection published a review⁸² 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 material, and some have specific options for CSEA, which eSafety regards as a key measure to ensure reports are actioned and prioritised appropriately.

eSafety asked about the ability of users to make reports of CSEA in-service in the context of sections 6 , section 11, section 13, section 14 and section 15 of the Determination.

Table 16: Ability for users to report in-service

In response to the Notices, the following information was given by providers regarding the ability to report in-service, (as opposed to navigating to a separate webform or email address).

⁸¹ Ofcom, ‘[Ofcom’s first year of video-sharing platform regulation](#)’, October 2022, accessed 11 September 2023.

⁸² Canadian Centre for Child Protection (2020) ‘[Reviewing child sexual abuse material reporting functions on popular platforms](#)’, accessed 11 September 2023.

Provider	Service	In-service user reporting option	Additional comments by provider
Google	YouTube Drive Meet Chat Google photos Blogger	Yes ⁸³	
	Google Messages Gmail	No	Users can report 'child endangerment' on Google services via a dedicated webform
Twitter	Public posts "tweets"	Yes	
	Direct Messages	No	Users can report CSEA through separate webform
TikTok	Photos/videos shared publicly (to everyone or followers only)	Yes	
	Direct messages TikTok live		
Twitch	Users signed into an account	Yes	
	Users not signed into an account	No	Twitch said it is working on a way for users who are not signed-in to be able to report. Users not signed-in can e-mail Twitch directly or contact Twitch Support on Twitter.
Discord	Direct messages	Yes	

⁸³ eSafety notes that for many of these services users must be signed into their Google account to make a report.

Provider	Service	In-service user reporting option	Additional comments by provider
	Server content	Yes	Some changes were made to enable in-service reporting to enable reports on web, desktop, and app after the end of the reporting period. Previously this functionality did not exist on all platforms.
	Server Livestreams	No	Reports can be made either for the server overall, or via a separate webform.
	Server audio	No	

7.13 Languages providers operate across

New CSEA material identified through machine learning tools, as well as harms such as grooming and livestreaming, typically require human moderator review to verify abusive material and activity flagged by technology or reported by users. Some harms, particularly grooming, are highly context dependent and require review of the surrounding conversation, as well as behavioural signals, in order to confirm. Many tools to identify harms like grooming only operate in English, or a small number of languages.

Beyond CSEA, other harms such as hate speech are particularly context dependent and require understanding of language and culture. It is therefore particularly important that companies have human moderators operating in the languages of the communities they offer services to. Relying on language translation tools risks losing important context and cultural or linguistic nuance.

The Australian Bureau of Statistics 2021 Census results list Mandarin, Arabic, Vietnamese, Cantonese and Punjabi as the top languages spoken in Australian homes, other than English. Not all providers operate across these 5 languages, let alone other major languages spoken in Australia and by the providers' wider users. This is important, as highlighted above, in instances of grooming and other harmful activities which can go undetected if moderators cannot confirm in the languages spoken by Australians.

eSafety asked about the languages providers operate across in the context of section 6, section 11, section 13, section 14 and section 15 of the Determination.

Table 17: Number and list of languages human moderators operate across

In response to the Notices, the following information was given by providers regarding the languages their human moderators operate across.

Provider	Number of languages	Languages
Google	'At least' 71	1. Afrikaans; 2. Albanian; 3. Amharic; 4. Arabic; 5. Assamese; 6. Azerbaijani; 7. Bahasa Indonesia; 8. Bahasa Malaysia; 9. Bangladeshi; 10. Bengali (Bangla); 11. Bihari; 12. Bulgarian; 13. Burmese; 14. Catalan; 15. Cebuano; 16. Chinese (Traditional and Simplified); 17. Croatian; 18. Czech; 19. Danish; 20. Dutch; 21. English; 22. Estonian; 23. Farsi; 24. Finnish; 25. French; 26. Galician; 27. German; 28. Greek; 29. Gujarati; 30. Hebrew; 31. Hindi; 32. Hungarian; 33. Icelandic; 34. Indonesian; 35. Italian; 36. Japanese; 37. Kannada; 38. Kashmiri; 39. Kazakh; 40. Khmer; 41. Korean; 42. Latvian (Lettish); 43. Lithuanian; 44. Malayalam; 45. Marathi; 46. Nepali; 47. Norwegian; 48. Oriya; 49. Persian; 50. Polish; 51. Portuguese; 52. Punjabi; 53. Quechua; 54. Romanian; 55. Russian; 56. Sanskrit; 57. Serbian; 58. Slovak; 59. Slovenian; 60. Spanish; 61. Swahili (Kiswahili); 62. Swedish; 63. Tagalog; 64. Tamil; 65. Telugu; 66. Thai; 67. Turkish; 68. Ukrainian; 69. Urdu; 70. Vietnamese. 71. Welsh
Twitter	12	1. Arabic; 2. Chinese; 3. English; 4. French; 5. German; 6. Hindi; 7. Indonesian; 8. Japanese; 9. Korean; 10. Portuguese; 11. Spanish, 12. Turkish Twitter stated that it 'has the ability to seek and conduct vendor services in a range of additional languages, which include but are not limited to those needed in the event of additional reviews, or for emergencies.'
TikTok	73 ⁸⁴	1. Afrikaans 2. Amharic 3. Arabic 4. Assamese 5. Azerbaijani 6. Bengali 7. Bhojpuri 8. Brazilian Portuguese 9. Bulgarian 10. Burmese 11. Cantonese 12. Croatian 13. Czech 14. Danish 15. Dutch 16. English 17. Estonian 18. European Portuguese 19. Filipino 20. Finnish 21. French 22. Garhwali 23. German 24. Greek 25. Gujarati 26. Haryanvi 27. Hausa 28. Hebrew 29. Hindi 30. Hungarian 31. Icelandic 32. Igbo 33. Indonesian 34. Italian 35. Japanese 36. Kannada 37. Kazakh 38. Khmer 39. Korean 40. Kurdish 41. Lao 42. Malay 43. Malayalam 44. Mandarin 45. Marathi 46. Minnan 47. Nepali 48. Norwegian 49. Odia 50.

⁸⁴ Following a subsequent question from eSafety, providing TikTok with a further opportunity to provide a list of the languages (rather than the total number), TikTok provided the list of languages.

Provider	Number of languages		Languages
			Persian 51. Polish 52. Punjabi 53. Rajasthani 54. Romanian 55. Russian 56. Serbian 57. Sesotho 58. Sinhalese 59. Somali 60. Spanish 61. Swahili 62. Swedish 63. Tamil 64. Telugu 65. Thai 66. Turkish 67. Twi 68. Ukrainian 69. Urdu 70. Vietnamese 71. Xhosa 72. Yoruba 73. Zulu
Twitch	Twitch Moderators	21	1. Arabic; 2. Chinese (Cantonese); 3. Chinese (Mandarin); 4. Chinese (Taiwanese); 5. Danish; 6. Dutch; 7. French; 8. German; 9. Indonesian; 10. Italian; 11. Japanese; 12. Korean; 13. Malay; 14. Norwegian; 15. Polish; 16. Portuguese; 17. Russian; 18. Spanish; 19. Swedish; 20. Thai; 21. Turkish
	Third-party vendor	19	1. Arabic; 2. Chinese (Cantonese); 3. Chinese (Mandarin); 4. Czech; 5. Dutch; 6. English; 7. French; 8. German; 9. Italian; 10. Japanese; 11. Korean; 12. Polish; 13. Portuguese; 14. Russian; 15. Spanish; 16. Swedish; 17. Thai; 18. Turkish; 19. Vietnamese Twitch noted that additional languages will and can be implemented when there is a need through additional third-party vendors.
Discord	29		1. German; 2. Korean; 3. Turkish; 4. Croatian; 5. French; 6. Polish; 7. Thai; 8. Czech; 9. Danish; 10. Russian; 11. Vietnamese; 12. English; 13. Dutch; 14. Spanish – Spain; 15. Traditional Chinese; 16. Finnish; 17. Italian; 18. Swedish; 19. Simplified Chinese; 20. Greek; 21. Japanese; 22. Brazilian Portuguese; 23. Bulgarian; 24. Hindi; 25. Lithuanian; 26. Norwegian; 27. Romanian; 28. Ukrainian; 29. Hungarian

7.14 Roles and responsibilities of volunteer moderators

Some service providers use volunteer moderators to support content moderation and enforce community rules. Volunteer moderators are given limited administrative power to remove material and ban accounts that breach community rules. Volunteer moderators are often appointed by the specific groups that they operate within.

There can be a lack of information on the division of responsibilities between volunteer moderator systems and some providers' professional trust and safety staff. eSafety asked providers what standards policies were available outlining the responsibilities of volunteer moderators, the action providers take to hold volunteer moderators to these standards, and the systems for escalating and communicating cases involving illegal material such as CSEA to professional trust and safety staff when identified by volunteers.

These questions relate to section 6, section 11, section 13, section 14, and section 15 of the Determination.

Table 18: Roles and responsibilities of volunteer moderators

In response to the Notices, the following information was given by providers.

Provider	Are professional trust and safety staff automatically notified of a CSEA violation?	Is there an enforcement standards policy outlining responsibilities/ expectations of volunteer moderators?	Can users directly report volunteer moderators for failing to enforce rules?	Action taken when a volunteer moderator fails to meet enforcement standards
Twitch	No	Yes	No	Twitch said that if a channel owner or volunteer moderator endorses or ignores violative content on their channel, then depending on the severity of the conduct, they can receive a warning, a suspension of varying lengths or indefinite suspension. Twitch says it suspended 7,312 creators for ‘unmoderated hateful conduct or harassment’ during the reporting period.
Discord	No	No	No	Discord said it doesn’t have dedicated avenues for complaints about volunteer administrators and moderators. Discord said it will take action against communities that violate its policies – including banning all members and shutting down the server.

8. Transparency summaries: Individual provider responses

8.1 Google Summary

Overview

Google LLC (Google) was asked about its YouTube, Drive, Meet, Chat, Google Photos, Google Messages, Gmail and Blogger services.

Note: eSafety used the term child sexual exploitation and abuse (CSEA) in the Notice given to Google. However, Google’s responses sometimes used the term CSAM. referring to ‘child sexual abuse material’, rather than the broader category of CSEA.

1. Questions about known and new CSEA material

Known CSEA images

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

Table A

Service	Hash matching to detect known images?	Name of tools used
Drive (when content is stored)	Yes (consumer only)	PhotoDNA SHA 256 Other proprietary technologies
Drive (when content is shared)	Yes (consumer only)	
Chat	Yes (consumer only)	
Google Photos	Yes	
Gmail	Yes (consumer only)	
Blogger	Yes	
Meet	No	
Google Messages	No	

Google stated that it does not use hash matching tools to detect known CSEA images on Google Messages given that in 'some instances' Google Messages includes end-to-end-encrypted (**E2EE**) communications, which Google stated makes detection 'more technically challenging' as well as being subject to greater privacy and security considerations. Google stated that it does not use hash matching tools to detect known CSEA images on Meet because 'Meet is a live video communication or conferencing tool that does not store images.'

In answer to what alternative reasonable steps Google is taking to detect known CSEA on the services where tools are not used, Google stated that any consumer account holder is able to report abuse, such as 'child endangerment', to Google within Meet. Google noted that users can also attach a video clip to provide context to the report. Google stated that 'child endangerment' reports are reviewed by a specialist team of child safety reviewers who take appropriate action. Guidance is also published by Google on this.

Known CSEA video

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

Table B

Service	Hash matching to detect known video?	Name of tools used
YouTube	Yes	CSAI Match for Video (Google's proprietary CSEA detection and removal technology)
Drive (stored content)	Yes (consumer only)	
Drive (shared content)	Yes (consumer only)	
Google Photos	Yes	
Blogger	Yes	
Meet	No	
Chat	No	
Google Messages	No	
Gmail	No	

Google stated that it does not use video hash-matching tools on Chat and Gmail because it is 'more challenging than image processing and detection'. Google noted that it continues to

evaluate deploying video hash-matching tools on these services. Google also stated that it deploys CSEA image hash-matching tools on both services and added that when an account is flagged for a CSEA violation on these services, the subsequent review process may uncover CSEA videos which will then be reported to the National Center for Missing & Exploited Children (**NCMEC**).

Google reiterated its stance that there are specific platform design challenges and heightened privacy and security considerations that ‘weigh against’ using automated tools to proactively scan for CSEA videos on Meet and Google Messages.

Sources of CSEA hashes

Google stated that it obtains hashes of known CSAM from ‘a variety of highly trusted sources’, including the Internet Watch Foundation (**IWF**); NCMEC; and ‘others’, including content that Google finds on its platforms and subsequently hashes.

Google noted that it reviews ‘CSAM hashes that have not been reviewed before independently by Google to confirm accuracy’.

Google stated that ‘Of the overwhelming majority of imagery reported by Google—approximately 90%-matches previously identified CSAM, much of which is already in the NCMEC database.’

In answer to a question about how often Google updates its hash lists, Google stated that it ‘continually’ updates its repository of hashes.

New or ‘unknown’ CSEA material

In response to questions about whether Google uses any tools to detect new CSEA material, Google responded that it uses tools on the following services:

Table C

Service	Tools used to detect new CSEA?	Names of tools used
YouTube	Yes	Machine Learning Classifiers (known externally as Google Content Safety API)
Drive (stored content)	Yes (consumer only)	
Drive (shared content)	Yes (consumer only)	
Google Photos	Yes	

Service	Tools used to detect new CSEA?	Names of tools used
Blogger	Yes	
Meet	No	
Chat	No	
Google Messages	No	
Gmail	No	

Google stated that on services that use these classifiers, human moderators review all reports flagged by its automated systems as suspected new CSEA material.

In response to why Google does not use automated tools to detect new CSEA material on Meet, Chat, Google Messages and Gmail, Google responded that it does not use classifiers to balance ‘the privacy interests of users’ and that it also takes into account ‘specific risk and abuse vectors’ relevant to each service.

Google also referred to technical and contractual constraints and ‘relevant privacy and security laws that apply’.

URL blocking

Google stated that it blocks URLs linking to known CSEA material on the following services:

Table D

Service	URLs linking to known CSEA material blocked?
YouTube	No
Drive (shared content)	
Drive (stored content)	
Meet	
Chat	
Google Photos	
Google Messages	
Gmail	
Blogger	

Google advised that it does not block URLs that link to known CSEA on any of the services covered by the Notice.

Google stated that it has strict policies that prohibit the use of its services to share or distribute CSEA material (including the sharing of links to known CSEA) and will remove that content which displays the URL and take other enforcement action as appropriate. If URLs linking to known CSEA content are identified, Google stated that it will de-link from Google Search,⁸⁵ report to NCMEC and ‘hash-match’ any CSEA content that appears.

Google submitted that, ‘Due to the changeable nature of content that may appear at a URL, hash-matching is an effective way of detecting and preventing the re-circulation of CSAM that may appear at those links.’

Proportion of CSEA detected proactively

Google provided the proportion of CSEA material it detected proactively during the Report Period on the following services, compared with material identified through user reports or trusted flaggers:

Table E

Service	Proportion of material detected proactively
YouTube	Approximately 99%
Drive (Consumer service)	
Chat (Consumer service)	
Google Photos	
Gmail (Consumer service)	
Blogger	‘More than’ 99% is caught as the result of ‘automated flagging’.

In its initial response to the Notice, Google did not provide a proportion for Meet and Google Messages, identifying the question as ‘not applicable’. Following a subsequent question from eSafety, seeking clarification of Google’s response to the Notice regarding Meet and Google Messages, Google responded that it defines ‘proactive detection’ as CSEA detected by its automated flagging systems, which it does not use on these services.

⁸⁵ Google Search was not captured by this Notice.

eSafety's understanding is therefore that 0% of material is detected proactively on these services.

Alternative steps taken to proactively detect CSEA

In response to questions about the alternative steps Google is taking to detect CSEA on its services, Google stated that it has 'strict policies' prohibiting CSEA on its services and that it 'makes extensive efforts to detect, remove, and report CSAM material' across all of its services. Google outlined several initiatives it has implemented to counter online CSEA on its services, including:

- Providing a webpage that provides information to people seeking to report CSEA to relevant authorities.
- Maintaining a discrete webform in the Google Help Centre that anyone can use to report 'child endangerment' on a Google service.
- 'Investing significantly' in its trust and safety teams, which Google stated operate 24 hours a day to respond to child safety incidents. Google stated that where they may be 'imminent harm to a child our specialist team escalates the report to NCMEC's prioritisation queue.'
- Acting as a 'lead member' of the Technology Coalition, an industry group that brings child safety experts, technologists and industry stakeholders together to collaborate on solutions to combat online CSEA.

Google stated that these broader initiatives often sit alongside the initiatives specific to particular services identified above.

2. Question about steps taken to identify whether an account that is found to have shared CSEA material contains other CSEA material

Google was asked whether it reviews an account for other CSEA material after it confirms that the account has shared CSEA material. If it does, Google was asked whether that review involves a review of all other content shared by that account, or a subset.

Google responded that if CSEA material is confirmed in connection with any Google Services accessed via the Google Account (which includes YouTube, Meet, Chat, Google Photos, Drive, Gmail, Blogger), Google 'may disable the account'. Google stated this will prevent use of any other Google services connected to that account to share CSEA material. Google may review a subset of other content in connection with that service for

some of the services identified. In answer to a question about whether the review of accounts for other CSEA material is carried out by human moderator and/or technical tools, Google stated that a combination of human moderator review and technical tools are used to identify CSEA content shared by violative accounts ‘depending on circumstance’.

3. Question about whether Google reviews content shared by an account on other Google services, when it is confirmed that that account has shared CSEA material

Google responded that it reviews an account’s activity across some other Google services when it confirms that CSEA has been shared by that account.

Google stated that when an account is found to have shared CSEA material, it ‘may be disabled’ thereby preventing it from being used to further distribute CSEA material on any associated services. Google highlighted that when an account has shared CSEA material on YouTube, Drive, Meet, Google Photos, Chat or Blogger, it may then review ‘other content in the account to identify additional pieces of CSAM’. Google stated that this review focusses on ‘products that are more likely to host or store user generated content’. Google also stated that ‘depending on the circumstances’ this review process may involve human moderators and the use of hash-matching and machine learning classifiers.

4. Questions about measures in place to detect the livestreaming of child sexual exploitation and abuse on YouTube

Google was asked if it had any measures in place to detect the livestreaming of CSEA on YouTube. Google responded that it does have measures in place and uses the following tools:

Table F

Measures in place to detect CSEA in livestreams or video calls/conferences		
Services	Yes/No	Names of tools used
YouTube	Yes	Machine learning language tools (audio and text) Video classifiers

Google was asked whether human moderators review all reports flagged by tools to confirm the livestreaming of CSEA. If not, Google was asked to specify the proportion that are reviewed by human moderators.

Google responded that human moderators review those reports ‘that are identified as having a high likelihood of being violative.’

Following a subsequent question from eSafety, seeking clarification of Google's response to the Notice regarding the proportion of reports flagged by tools, that are sent to human moderators, Google responded:

Human moderators do not review all content that is flagged. We rely on various signals to determine whether human moderators will review a video that has been flagged. Human moderators review those that are identified as having a high likelihood of being violative.

5. Questions about language analysis technology used to detect CSEA activity

Language analysis technology to detect grooming

Google reported that it uses machine learning classifiers to detect likely online grooming on YouTube. Google stated that its human moderators review a subset of the total reports flagged by this system.

Google stated that it does not use any language analysis technologies to detect likely online grooming on Meet, Chat, Google Messages, or Gmail. Google explained that this is due to ‘heightened privacy and legal considerations’ which apply to using language-based analysis tools on ‘private electronic communications’ as compared to YouTube.

When asked to explain the alternative steps it is taking to protect children from being groomed on these services, Google responded that there are design constraints on Meet, Chat, Google Messages and Gmail that prevent users messaging another user without knowing the user’s address, the presence of blocking functions, and information that helps users report CSEA to Google and relevant authorities.

Google also referred to its Family Link service, which enables parents to create and supervise an account on behalf of their child.

Google also noted the following additional measures on YouTube:

- The default upload settings for users aged 13 to 17 is set at “private”.

- Content that doesn't violate Google policies but features children may have some features disabled at both the channel and video level.
- With a supervised account, parents can limit the videos and music children under 13 can find and play, and the features that can be used (including comments, handles and live-chats).

Language analysis technology to detect terms, abbreviations, codes and hashtags indicating likely CSEA activity

Google answered 'yes', in answer to a question about whether YouTube uses language analysis technology to detect likely terms, abbreviations, codes and hashtags indicating likely CSEA activity, in particular but not limited to sexual extortion and the exchange of gifts for underage sexual acts.

Google answered 'no' for Meet, Chat, Google Messages, and Gmail for the reasons set out in section 5 above.

Google reported that YouTube uses machine learning classifiers to detect comments that 'may indicate CSEA activity'. Google added that it doesn't match specific words and abbreviations against a known list of violative terms on YouTube, because it stated that the terms used by offenders evolve. Google explained that YouTube's machine learning models are 'trained to understand contextually what is being said, so as to more effectively detect CSEA activity.'

6. Question about information collected from users on sign up

Information collected

When asked about information it collects when a user signs up to a service, Google responded that a user can access YouTube, Drive, Meet, Chat, Google Photos, Google Messages, Gmail and Blogger using their Google account.

Google added that the following information is captured on creation of a Google account:

- Name
- Email
- Birthdate
- User can choose to provide a phone number

- User can choose to provide a secondary email or link to a payment profile
- IP address (at the point of account creation or acceptance of Google's Terms of Service)
- Device Identifiers (at the point of account creation or acceptance of Google's Terms of Service)

Google reported that some Meet legacy users can sign in to Meet with their phone number only, and Google Messages account users can also use Google Messages solely with their phone number.

Google noted that additional information may also be captured on sign up for the following services:

- YouTube – payment information may be collected if the user subscribes to a paid YouTube product or service or is a monetising creator.
- Drive, Meet, Chat, Gmail – additional information may be collected for paid services, for example Google Workspaces
- Blogger – Google also collects the IP address at the time of Blog creation, but this data is anonymized after 240 days.

Authentication of accounts

In answer to a question about whether Google has an authentication process for any of the information captured on sign up across its services, Google reported that 'in certain circumstances' a user may be asked to verify the information provided, for example when they add payment information or use a non-Google e-mail to set up a Google account. Google also noted that for Google Messages, it will periodically verify a user's number with the relevant carrier to ensure the number is still active.

Google explained that it applies additional verification aimed at detecting bots, spam and abusive behaviour. In such situations, Google may require users to verify their account via verification sent via text message.

7. Questions about steps to prevent recidivism

Google listed a minimal⁸⁶ number of indicators used to detect users on Drive, Meet, Chat, Google Photos, Gmail, and Blogger that have previously been banned for CSEA-related breaches, which eSafety has decided not to publish to avoid the information being misused.

A minimal number of indicators are also used on YouTube. However, YouTube requires additional verification to unlock some features, such as livestreaming. A higher level of verification is also required for embedding livestreams, monetisation, and making community posts. In order to access these features, Google stated that a user must either establish sufficient channel history on the platform or verify their identity by providing a government ID or a short video of the person's face. Google stated that this information is used to identify repeat offenders.

Google stated in response to the Notice that it does not use any indicators to address recidivism on Google Messages for users that have not signed up via a Google account. It subsequently corrected this answer. Google Messages uses a minimal number of indicators.

When asked why any indicators it captures on sign up to Google services (as set out in section 6) are not used to aid in addressing recidivism on its services, Google replied that it:

invests significant resources in detecting, removing and reporting instances of CSAM on its platform, and collaborating with industry and NGO partners to prevent re-circulation of material across the network.

More broadly, Google stated that when it confirms 'CSAM' in a user's account, the material will be removed and the account 'may be disabled'.

Google added that it will notify the user and explain that their account has been disabled due to a 'CSAM-related' violation. Google stated that if a user's account is terminated for a CSAM violation, the account will lose access to all Google services and products that are 'based on or dependent' on that Google account.

When 'CSAM' is detected, Google stated that content is reported to NCMEC, which liaises with law enforcement agencies. In the period from January to June 2022, Google noted that

⁸⁶ eSafety uses the following terms to give an impression of the extent of the indicators used by services in the table below, rather than publishing the specific indicators which could be misused:

- Minimal: A small number
- Several: A moderate number
- Multiple: A significant number.

Google and YouTube combined reported a total of 6,698,201 total pieces of content to NCMEC.

8. Questions about user reporting in relation to CSEA material and activity

‘In-service’ reporting

Google was asked if users can report instances of CSEA activity or material to Google within a service (as opposed to navigating to a separate webform or e-mail address). Google provided the following information for the services covered by the Notice:

Table G

Service	Users can make in-service reports of CSEA
YouTube	Yes
Drive	
Meet	
Chat	
Blogger	
Google Photos	
Google Messages	No
Gmail	

Google explained that where in-service reporting is not available, users can report ‘child endangerment’ on Google services via a dedicated webform. eSafety notes that this webform requires a user to log in with a Google Account.

eSafety also notes that users who are not signed into a Google Account on YouTube cannot make an in-service report.

Median time for a user report of CSEA to be actioned

Google was asked to provide the median time to respond⁸⁷ to a user report of CSEA for YouTube, Drive, Meet, Chat, Google Photos, Gmail, Blogger and Google Messages.

Table H

Provider	Services	Median time for user reported CSEA material to be actioned
Google	YouTube Drive Meet Chat Google Photos Gmail Blogger Google Messages	<p>Google did not provide an answer for the median time from the point when a user makes a report to the report being actioned.</p> <p>Instead Google stated that:</p> <p>‘to the extent we are able to calculate this per product, the median time it takes Google’s Child Safety team to action that content once it is flagged to the Child Safety team as CSAM is between 30-60 minutes. [This].. represents response time for user reports from Australia for the reporting period.’</p> <p>As outlined in Table I Google also provided the percentage of content that violates YouTube’s Child Safety policy (which includes but is broader than CSEA) with 10 or fewer views at the time of removal.</p>

Google also stated that:

If a user is aware of CSAM on Google’s platform, our Help Centre directs users first and foremost to report this material to the appropriate organisation or to the police. The Help Centre directs users to submit CSAM to US NCMEC, which can direct reports to law enforcement around the world, and also provides contact information for local organisations around the world, including the Office of the eSafety Commissioner in Australia. If a user suspects that a child is in immediate danger in any way, Google directs that user to contact the police immediately.

In turn, Google partners with NCMEC and other NGOs to receive reports of CSAM on our platform. When reports are received from these NGOs, we prioritise and take action on that content as soon as possible.

We also allow users to submit reports of child endangerment, inappropriate content, or inappropriate behavior towards children on specific Google products. Again, if a user is aware of immediate danger, exploitation, or trafficking of a child, we direct the user to contact the police immediately.

⁸⁷ Defined in the Notice as ‘Content removed and reported, user banned, or other content moderation decision made.’

eSafety follow-up questions seeking median time to respond

eSafety followed up providing Google with a further opportunity to provide the information sought in the Notice, requesting both service specific information, and also the *overall response time from the time the user reports*, not the time to respond once allocated to an internal specialist team.

Google reiterated that when content on Meet, Drive, Google Photos, Blogger, Gmail, and Chat is reported as CSEA by a user, it estimates that the median time taken for its *Child Safety Team* to action that content is 30 – 60 minutes. Google did not provide any figure for the time to respond from the point at which a user had made the report, as required. No information was provided in response to eSafety's follow-up question in relation to Google Messages.

In its initial response to the Notice Google also highlighted that the 30 – 60 minute figure represented response time for 'user reports from Australia for the reporting period'. In response to eSafety's follow-up questions, Google revised this to say that the 30 – 60 minutes is an 'estimate based on global reports and is not Australian specific'.

Google also stated that it is 'complicated' to calculate the median time due to various factors including: users having multiple different pathways to flag CSEA material; users potentially incorrectly flagging content as CSEA material; and because Google's Child Safety Team does not 'prioritise complaints received for one service over any of the other services listed'.

Google stated that it was unable to provide a median response time for YouTube because it did not have 'readily available data'. Google stated that calculating median response time to user reports of CSEA on YouTube was complicated by 'a number of factors'. Google cited the lack of a CSEA-specific in-service reporting function (for example its 'child abuse' category could include other forms of abuse), the fact that videos could be flagged multiple times and for different reasons, content may be removed for a violation of YouTube's Child Safety Policy that was not originally identified or reported as CSEA, and that YouTube's Child Safety Policy is broader than just prohibiting CSEA. Google also stated that approximately 99% of CSEA material on YouTube is proactively detected and removed by YouTube's automated flagging systems.

Instead, Google provided the percentage of videos that violated YouTube's Child Safety policy (which includes types of harmful content other than just CSEA such as (amongst others) dangerous acts involving minors, infliction of emotional distress on minors and cyberbullying or harassment involving minors) and were removed before they received more

than 10 views. Google suggested that measuring the number of ‘views’ before violating content is removed is an effective measure of YouTube’s responsiveness in detecting and removing violating content on YouTube.

Table I

Percentage of videos removed in Australia for violation of YouTube’s Child Safety Policy that had 10 or fewer views at time of removal	
2022 Q1	Over 80%
2022 Q2	Over 80%
2022 Q3	Over 80%
2022 Q4	‘About’ 90%

9. Question about the languages that Google’s human moderators operate across

Google was asked to list all languages that its human moderators (employees and contractors) operate across.

Google stated that ‘review of CSAM material is typically an image-based review’.⁸⁸ Google stated that its moderators review CSAM material in English and ‘have support for’ 17 additional languages:

Table J

Arabic	Bahasa Melayu	Bahasa Indonesia	Cantonese
Croatian	Czech	German	Greek
Italian	Korean	Mandarin	Polish
Portuguese	Russian	Spanish	Tamil
Turkish			

After Google’s response to the Notice, eSafety followed-up to clarify whether these were all the languages that Google’s human moderators operate across, or if there were additional languages used in the review of material and activity other than CSEA. Google

⁸⁸ eSafety notes that review of CSEA activity such as grooming is highly contextual, and not limited to an ‘image-based review’.

clarified that its ‘language capability may vary at any given time’ and that while it ‘cannot provide a comprehensive list of all language capability for human moderators’ it operates across ‘at least’ the following languages:

Table K

Afrikaans	Albanian	Amharic	Arabic	Assamese	Azerbaijani	Bahasa Indonesia	Bahasa Malaysia	Bangladeshi
Bengali (Bangla)	Bihari	Bulgarian	Burmese	Catalan	Cebuano	Chinese (Traditional and Simplified)	Croatian	Czech
Danish	Dutch	English	Estonian	Farsi	Finnish	French	Galician	German
Greek	Gujarati	Hebrew	Hindi	Hungarian	Icelandic	Indonesian	Italian	Japanese
Kannada	Kashmiri	Kazakh	Khmer	Korean	Latvian (Lettish)	Lithuanian	Malayalam	Marathi
Nepali	Norwegian	Oriya	Persian	Polish	Portuguese	Punjabi	Quechua	Romanian
Russian	Sanskrit	Serbian	Slovak	Slovenian	Spanish	Swahili (Kiswahili)	Swedish	Tagalog
Tamil	Telugu	Thai	Turkish	Ukrainian	Urdu	Vietnamese	Welsh	

10. Question about information provided in Google’s reports to the National Center for Missing and Exploited Children (NCMEC)

In response to a question about the information Google includes in its reports to NCMEC on specific services, Google originally responded by providing examples of the information reported to NCMEC across all its services, rather than answering for each specific service as required.

Following a subsequent question from eSafety, providing Google with a further opportunity to provide the information sought in the Notice, Google provided a breakdown of information that it provides in a NCMEC report for each service as required. eSafety has correlated this with the types of information NCMEC defines as ‘actionable’ for law

enforcement. This includes information commonly provided across all Google products, as well as information specific to particular services covered by the Notice.

11. Questions about end-to-end encryption (E2EE) on Google services

All interventions that Google uses to detect CSEA in E2EE Google Messages

Google was asked to provide details of all interventions it has put in place to detect CSEA on Google Message's Rich Communications Service (RCS), on which E2EE is implemented.

Google responded that:

Tools to intervene in encrypted spaces include, among other things, responding to user reports and using behavioural information and meta-data signals, which can be deployed to detect behaviours that may be putting children at risk.

eSafety followed up with Google after it had responded to the Notice, noting that Google's response did not answer the question, as it did not confirm what interventions are in place to detect CSEA, if any. eSafety provided Google with a further opportunity to provide the information sought in the Notice. Google responded that:

There is no tool used to proactively detect CSEA content in RCS messages that are end-to-end encrypted. Users of Messages are able to send feedback to Google, attaching screenshots and/or system logs as appropriate. Google is then able to assess user reports and relevant metadata signals to respond appropriately.

Google also referred to other data or signals (metadata), which it said can be shared with law enforcement when provided with valid law enforcement requests.

All steps Google has taken to investigate means of detecting CSEA within E2EE services

Google was asked to provide information on all the steps it took during the Report Period to investigate means of detecting CSEA within E2EE services. Google was also asked to list all research or initiatives to detect CSEA within E2EE services, their status, and any decisions made during the Report Period to implement them.

In Google's response to the Notice, it referred broadly to the work of the Technology Coalition – a tech industry association that facilitates inter-industry collaboration on technical solutions to counter online CSEA – but did not provide any information on steps taken by Google.

eSafety followed up with Google noting that Google had not provided the information required, and providing Google with a further opportunity to provide the information sought by the Notice.

In response, Google referred to its April 2023 publication of a research paper on ‘novel protocols for on-device blocklisting’. The paper, ‘*Robust, privacy-preserving, transparent and auditable on-device blocklisting*’,⁸⁹ was described by Google as involving research into methods ‘that allow a client to determine if an object is harmful based on threat information possessed by a so-called remote enforcer in a way that is both privacy-preserving and trustworthy’. Google added that it considers that publishing this research represents ‘a critical step towards enabling complex, multi-stakeholder discussions on how best to provide on-device protections against abuse in E2EE environments.’

eSafety notes that it is not clear from Google’s subsequent response whether this constitutes the totality of the research or initiatives to detect CSEA within E2EE services undertaken by Google during the Report Period.

Work undertaken to implement E2EE on its services

Google was asked to list any consumer services (or parts of services) on which it had carried out work during the Report Period with the goal of implementing E2EE.

Google responded noting:

The fight against CSAM is a difficult one and perpetrators are sophisticated and work hard to constantly bypass the systems. This means that we are constantly investing in updating our technology to ensure that it continues to be precise and effective. While Google remains committed to continuing to develop new products and features in ways that help keep children safe and at the same time preserve user security and privacy, it does not comment on future product plans.

eSafety followed up with Google noting it had not provided the information required about work conducted during the Report Period. eSafety provided Google with another opportunity to respond and provide the information sought by the Notice. In response, Google stated:

⁸⁹ Cornell University (2023), ‘[Robust, privacy-preserving, transparent, and auditable o-device blocklisting](#)’, accessed 9 October 2023.

Google has a large number of products and services, and is continually working to improve existing products and to launch new products. It is not feasible to list or confirm for every consumer service or product (including for new products or features still in development and not yet announced) on whether any work may have been carried out for the potential use of encryption technologies (including E2EE or equivalent) during the Reporting Period.

Google added that apart from Google Messages,

we confirm that there were no publicly announced changes to the use of E2EE for these consumer products or services [the services covered by the Notice] during the Reporting Period.

eSafety notes that the Notice did not ask for information on ‘publicly announced changes’.

12. Questions about steps to identify and remove underage users

Tools to detect underage users

Google stated that it requires users to provide their date of birth when they register a Google account and if this indicates they are under 13 years old they will be prevented from registering without ‘parental approval and supervision of their account until they are an adult’. Google added that if a user self-declares that they are between 13 and 17, they will not be able to access age-restricted content until they turn 18.

In addition to age-gating on sign-up, Google reported that it also uses machine learning systems to identify underage users that have entered a false birthdate to access age-restricted content. Google explained that it also deploys machine learning systems to assess user’s activity to confirm that they are an adult and can access mature apps or videos. eSafety has chosen not to publish the specific factors Google reported that its model considers to prevent this information being misused. Google noted that when its model ‘fails to substantiate that a user is an adult’, that user will need to provide further verification of their age, such as a photograph of government ID or allowing an authorisation on their credit card.

Google stated that YouTube also uses classifiers to search for ‘content signals’ indicating that an account is operated by someone under the age of 13. Google added that YouTube also uses classifiers to determine whether ‘young minors’ are livestreaming themselves

without adult supervision. When these classifiers flag an account, Google added that it is sent to human moderators for assessment.

Number of underage users identified

Google was asked to provide information about the number of users identified as underage on YouTube, Drive, Meet, Chat, Google Photos, Google Messages, Gmail and Blogger services as well as information about underage users removed globally and in Australia. It was also asked to provide the following information:

- 1) The number of underage users removed globally and:
 - i. The global proportion of underage users detected proactively
 - ii. The global proportion of underage users reported by users
- 2) The number of underage users removed in Australia and:
 - i. The Australian proportion of underage users detected proactively
 - ii. The Australian proportion of underage users reported by users

Google's response to the Notice did not provide the answers for each service in the template provided to it. It provided an answer instead that stated that 19 million accounts across all Google products globally had been detected or flagged as potentially operated by someone underage. Google stated that the 19 million accounts identified referred to:

the total global number of accounts that were detected or flagged by our systems as potentially operated by someone that is underage. If an account holder is flagged as "underage," they will be asked to provide a government ID or credit card to verify their age or to add parental supervision to their account through Family Link.

We do not have data for the number of accounts held by underage users (as defined by the location of the user, which for Australia is 13) reported by other users. We also do not have Australian-specific data for the reporting period.

Following a subsequent question from eSafety, providing Google with a further opportunity to provide the information sought in the Notice regarding information on the individual services, Google stated that it was not able to provide a breakdown for each service because:

A user accesses Gmail, Meet, Chat, Drive, Photos and Blogger by logging into their Google Account, and must be logged in to their Google Account to access certain YouTube features such as creating videos, “liking” videos, subscribing to channels and creating a YouTube channel. A user does not create a separate account to use each of these services.

Our handling of underage accounts is not product- or service-specific, but impacts the user’s Google Account as a whole, and therefore their use of all services is impacted once they are identified. For services that can be accessed without a Google account, such as Messages which can be used via a User’s mobile phone number, we do not hold information relevant to this request.

13. Questions about protections for minors on YouTube

YouTube’s website states that the service disables comments by default on videos that are ‘made for kids’ or feature ‘minors’ ‘that could be at risk of attracting predatory behaviour’. Google was asked to explain how it applies these restrictions. Google stated that in addition to requiring YouTube ‘creators’ to self-designate that their videos are ‘made for kids’, YouTube also uses a machine learning classifier that assesses ‘a number of risk factors’ to determine whether content ‘is not policy violating but is at risk of predatory behaviour’.

eSafety has chosen not to publish the specific signals used by this classifier to prevent the information being misused.

14. Questions about Google recommender systems

Percentage of content watched on YouTube that is algorithmically recommended

In the Notice eSafety noted that in 2018 YouTube’s Chief Product Officer stated that 70% of content watched on YouTube was recommended by YouTube’s algorithm, and asked Google for an updated figure for the Report Period.

Google replied that it did ‘not have an updated or a revised figure for the reporting period.’

eSafety followed up with Google to seek clarification of Google’s response to the Notice, and asked whether Google held any other data or information to answer the question for any part of the Report Period.

Google responded that it was not ‘in the time available to respond to the original notice, able to calculate a revised or updated percentage of content watched on YouTube that was recommended by YouTube’s algorithm.’

While Google provided a different metric, eSafety agrees that this alternative metric cannot be compared to the 2018 figure quoted above, so has not been included in this summary.

eSafety notes that Google had the opportunity to seek extensions to the Notice period, but did not seek one.

Design objectives of YouTube’s recommender systems

Google was asked to describe the design objectives of the systems it uses to recommend content on YouTube. Google pointed to information it had previously made publicly available on YouTube’s official blog page. This included the statement that YouTube’s recommendation system is ‘built on the simple principle of helping people find the videos they want to watch and that will give them value’ and that it does so by comparing ‘a person’s viewing habits with those that are similar and uses that information to suggest other content’ that they may want to watch.

Google also stated that it uses classifiers to ‘minimise chances that a viewer will see problematic content and connect viewers with high quality information.’ The classifiers identify whether a video is ‘authoritative’ or ‘borderline’ for content that comes close but doesn’t violate Community Guidelines. Google pointed to publicly available guidelines for supporting human evaluators who support the classification of content.

Signals used to recommend content on YouTube

Google was asked to list and describe the primary signals it uses to determine what content should be recommended to a user and to provide details of their relative importance. Google stated that it uses 80 billion pieces of information every day to help inform its system about what a user finds satisfying and stated that it was not possible to list the top 10 signals.

However, Google did identify 4 signals its system considers:

- Clicks
- Watch time

- Survey responses
- Sharing, likes, and dislikes

Google reported that the importance of a particular signal will depend on the individual user and that their system ‘doesn’t follow a set formula, but develops dynamically as [users’] viewing habits change’. Google added that signals can be overruled in the interests of meeting ‘our responsibility to the YouTube community and to society’.

Measures to test and update YouTube’s safety systems

Google was asked to describe the measures YouTube employs to test and update its systems, in a way which contributes to the overall safety of YouTube and avoids the risk of amplifying harmful content.

Google stated that to ensure the overall safety of the service and avoid the risk of amplifying harmful content, YouTube uses a metric it calls ‘violative view rate’, which involves human reviewers sampling user views of YouTube videos and assessing the videos to calculate an estimated proportion of the total views that were of videos that violated its Community Guidelines. YouTube stated that the violative view rate for the third quarter 2022 was 0.10 – 0.11%. Google also highlighted that it compares the number of appeals submitted by creators in response to videos that are removed. Google calculates these metrics quarterly and publishes the information in its transparency report.

eSafety followed up with Google to clarify its response to the Notice, and asked Google to clarify if tracking appeals and violative view rate metrics were the only measures YouTube employs to test and update its recommendation system to improve safety on the service. Google responded by stating that the ‘YouTube Recommendation System is complex’ and that ‘engineers continually adjust and test changes to search and recommendation which, along with user feedback and other signals, are used to update and improve recommendations over time.’

Google referenced other metrics it publishes in its transparency report such as number of videos, channels, comments removed, and a separate report it publishes on enforcement of policies against violent extremism, hate-speech and child safety. Google also stated that since 2019, it has adjusted its system to ‘prioritise high quality content in ranking recommendations and not widely recommend ‘borderline content.’ Google defines ‘borderline content’ as ‘content that comes close, but doesn’t violate [YouTube’s] Community Guidelines’. Google reported that its aim is to keep consumption of borderline content to ‘significantly below 1%’.

Google also outlined several ongoing initiatives to review and develop YouTube's Community Guidelines, including:

- The YouTube Intelligence Desk which specialises in identifying new potential violative trends
- Testing of enforcement guidelines by content moderators
- Regular meetings of YouTube's trust and safety specialists to assess the way individual policies are enforced.

Criteria used to evaluate content that is harmful to children

Google was asked to describe the criteria YouTube uses to evaluate content that is harmful to children but does not violate community guidelines for the purposes of preventing children from being recommended this kind of content. Google's response consisted of information it has previously made publicly available on its YouTube Help webpage.

Google stated that it may apply age-restrictions to content that doesn't violate its policies but is not appropriate for viewers under 18. Google added that this restriction can apply to 'videos, video descriptions, custom thumbnails, live streams, and any other YouTube product or feature' and it will prevent viewers who are under 18 or who have not logged into an account from accessing the relevant content.

Google noted that it will consider applying age-restriction to content that contains one or more restricted themes:

- Child safety; including videos of adults participating in dangerous activities that minors could easily imitate
- Harmful and dangerous activities; including regulated substances and drugs
- Nudity and sexually suggestive content
- Violent or graphic content
- Vulgar language

Google also referenced its Quality Principles for Kids and Family Content which it says it uses to guide the way it surfaces content in YouTube's recommendations.

15. Question about the steps YouTube takes to prevent users sexualising content involving children

Google was asked to explain the steps YouTube takes to prevent users from sexualising content such as innocent videos posted by children of them dancing, singing or unboxing toys. Google referred to publicly accessible information in its transparency report that it removed 2,017,231 videos for violating child safety policies (though these were not all CSEA) in the period July to September 2022, of which 99% were detected by automated systems. Google also specified the following steps YouTube has in place to prevent users from sexualising content involving children:

- Prohibitions against content that endangers the emotional and physical well-being of minors, including sexually explicit content featuring minors or content that sexually exploits minors.
- Uploading, streaming, commenting, or engaging in activity that harms minors results in the content being removed and the account may be terminated.
- Reporting child sexual abuse imagery to NCMEC.
- Disabling comments, live chat, live streaming, video recommendations and community posts on content that features minors. Content designated as ‘made for kids’ has comments disabled by default.
- Content that is uploaded by minors (13-17), has video visibility settings for creators aged 13-17 set at the most private option available, by default.

16. Question about the steps taken to prevent autocompleting of YouTube searches recommending terms associated with CSEA

To prevent the autocomplete of YouTube searches recommending phrases that are likely to be associated with CSEA, Google says YouTube has systems in place designed to prevent policy-violating autocomplete predictions. Users can also report policy violating autocomplete predictions at the bottom of the prediction box.

17. Questions about the use of metrics to internally assess efficacy of interventions to detect CSEA

Google was asked if its services have internal metrics in place to assess the efficacy of its interventions to detect CSEA material. Google responded by directing eSafety to the information it provides in its transparency report.

eSafety notes that it sought information on Google's internal metrics for assessing efficacy of its interventions to detect CSEA. While Google pointed to its transparency reports which it stated contains internal metrics, it is not clear to eSafety whether Google has additional internal metrics which are not included in its public transparency report.

18. Question about appeals against CSEA-related moderation

Google was asked to provide the number of appeals made by users during the Report Period for content removed or accounts banned for CSEA, and the number of appeals that were successful for its YouTube, Drive, Meet, Chat, and Google Photos services. Google stated that it did not have data for 'all applicable products'.

Google reported:

While Google does not currently have data for all applicable products, reinstated accounts are a tiny fraction of all disabled accounts and a small amount of accounts seeking appeal.

Google added that:

We also note that as part of our global transparency reporting, we publish the total number of accounts disabled globally for each half year period.

When we confirm CSAM in a user's account, that content will be removed and the account may be disabled. When we disable an account after finding child sexual abuse content, we notify the user that the account has been disabled due to content having been found that involves a child being sexually abused or exploited. Our message clearly states that this is a violation of our policies and may be illegal. The message also links to our disabled account help centre page, which describes in more detail why accounts are disabled for child sexual abuse and exploitation, informing users that we will report child exploitation to NCMEC. Our communication gives the user a link to start the appeals process if they feel the account was disabled by mistake. Our child safety team reviews the appeal, and we reinstate the account as soon as possible if we find we have made a mistake. Our reinstatement rate based on content-level false positives (e.g. the imagery detected was determined not to be CSAM) is extremely low.

eSafety followed up with Google seeking clarification of its response to the Notice. eSafety asked Google to confirm whether it held any appeals data for the services listed as requested in the Notice.

Google responded that it was not able, in the time available to respond to the original Notice, to provide the number of appeals and of those, the number of appeals that were successful for each of the listed services for a CSEA violation. However, with the additional time, it had been able to collect additional information set out below.

eSafety notes that Google had the ability to seek extensions to the Notice but did not seek one.

In response to eSafety’s follow-up question, Google provided the following additional information for the period 1 May 2022 to 31 March 2023:

Table L

Number of appeals of a decision to disable a Google Account held by an Australian user for CSAM:	407
Of those appeals, the number of Google Accounts reinstated:	15
Google emphasised that these figures should not be interpreted as equivalent to an ‘error rate’, as there are a number of reasons as to why an account may be reinstated.	

Google stated that it provided the number of appeals and reinstatements for the period 1 May 2022 to 31 March 2023 because it did not ‘hold/maintain data for the entire requested Reporting Period.’

Google added that it provided the number of Australian users⁹⁰ who appealed a decision to terminate a Google Account (which may include services other than the listed services), rather than the number of appeals for each listed service, for the following reasons:

⁹⁰ Google said it determines that a user is Australian ‘largely by either the country from which the user created their account or the country from which the user most often accesses Google services.’

When we refer to disablement of a “user’s account” for CSAM we are referring to disablement of a user’s Google Account. A user does not create a separate account for these services. If a child sexual abuse violation is detected in connection with any of YouTube, Drive, Meet, Chat and Photos, the User’s Google Account is what would be disabled, and if that user appeals and is successful, their Google Account will be reinstated.

‘The disablement of a user’s Google Account (and subsequent appeal and/or reinstatement) may be the result of CSAM detected on more than one service. For example, a user may have stored an image on Drive and Photos for the same account. In that case, there is one appeal and one reinstatement, not two separate appeals and two reinstatements for each separate service.

19. Questions about Safety by Design training for staff

Google stated that while it doesn’t provide training specifically focused on eSafety’s Safety by Design Principles, ‘equivalent principles are nonetheless incorporated into our existing internal processes and systems’. Google added that this includes:

- assessing early in the product design stage to assess any potential vectors for abuse, including potential child abuse and user security (amongst other risks);
- having in place policies governing acceptable use of Google’s products and services, and extensive infrastructure to investigate reports of abuse and take appropriate action in accordance with its policies;
- providing training for our enforcement teams on how to review and action policy violations (including child safety violations); and
- publishing transparency reports, which report on (amongst other things) combating CSAM, enforcement of YouTube Community Guidelines and government requests.

8.2 Twitter Summary

Overview

Twitter, Inc. was asked about its Twitter service. In March 2023, Twitter, Inc. merged into X Corp. In July 2023, the Twitter service was rebranded as X. The name ‘Twitter’ is used in this summary, as that was the name of the service at all points during the Report Period.

For some questions, Twitter was asked to provide information and data for certain periods of time within the Report Period. These questions were asked given Twitter’s change of ownership in October 2022, which has resulted in changes to Twitter’s staffing, as well as processes and procedures. eSafety’s questions aimed to identify whether these changes at Twitter had had an impact on safety on the service and Twitter’s implementation of the Basic Online Safety Expectations.

Twitter was provided with an advance draft of this summary in July 2023 prior to publication. It responded and stated that ‘it maintains a zero tolerance approach to combating child sexual exploitation (CSE) on its service and that any content that features or promotes child sexual exploitation (CSE) on its service is expressly prohibited.’ Twitter added:

whilst there were changes in organizational structure and staffing following the change of ownership in October 2022, [Twitter’s] policies regarding CSE, as well as the processes and procedures in place in respect of the same, remained the same.

Twitter outlined that any evaluation of the impact of the change of ownership on its efforts to address CSE should take the metrics it had published on its website into consideration.⁹¹

Note: eSafety used the term child sexual exploitation and abuse (CSEA) in the Notice given to Twitter. However, Twitter’s responses sometimes used the term CSAM (child sexual abuse material) or CSE (child sexual exploitation), rather than CSEA. Twitter clarified that it uses ‘an umbrella definition’ of ‘Child Sexual Exploitation’ (CSE) in its policies. Twitter stated that it considers that ‘CSE covers any material and/or behaviours which are sexually exploitative, as well as material that would be deemed to be sexually abusive (which encompasses sexual assault against a child) (“CSAM”)’.

⁹¹ [An update on Twitter Transparency Reporting](#), accessed 9 October 2023 and [Feb 2, 2023 post](#), accessed 9 October 2023

1. Questions about known and new CSEA material

Known CSEA images

In response to questions about Twitter's use of hash matching to detect known CSEA material, Twitter stated that it uses hash matching tools to detect known CSEA images on public posts ('tweets') and direct messages.

Table A

Service	Hash matching to detect known images?	Name of tools used
Tweets	Yes	PhotoDNA Safer by Thorn ⁹² Internal tools including ProfileViewer2, as well as internal machine learning models and classifiers that do not have names ⁹³
Direct messages	Yes	PhotoDNA, Safer by Thorn 'Other proprietary technologies', which Twitter subsequently named as ProfileViewer2 as well as internal machine learning models and classifiers that do not have names.

Twitter stated that it has a number of sources for hashes, including from the National Center for Missing and Exploited Children (**NCMEC**), the Internet Watch Foundation (**IWF**) and 'other partners'. Twitter stated that it takes all available hashes of CSEA images from these databases, and that it updates its hash lists daily.

Following a subsequent question from eSafety, providing Twitter with a further opportunity to provide the information sought in the Notice, which asked for all databases, Twitter added that it takes all hashes made available by Thorn Safer (including IWF, NCMEC and

⁹²Following a subsequent question from eSafety, providing Twitter with a further opportunity to provide the information sought in the Notice, Twitter provided the name of this tool.

⁹³Following a subsequent question from eSafety, providing Twitter with a further opportunity to provide the information sought in the Notice, Twitter provided this information. eSafety notes that these are not hash matching tools.

Canada’s Cybertip). Twitter added that this list may not be exhaustive as some organisations provide inputs to others.

Following this response, Twitter subsequently provided eSafety with further information, stating:

[Twitter] also has access to the ‘Exploitative Hash-Sharing’ database via Safer/Thorn. [Twitter] currently does not proactively hash against this database as it does not include NCMEC reportable content.. [Twitter] is focused on prioritizing our proactive enforcement against content that meets industry reporting requirements to NCMEC.

Known CSEA videos

In answer to a question about whether Twitter uses tools to detect known CSEA video, Twitter provided the following information.

Table B

Service	Tools used to detect known video
Public posts (‘tweets’)	No
Direct messages	No

Twitter stated that it does not use any tools to detect known CSEA video on public posts or direct messages because Twitter has ‘been developing the technology needed to support this and will be launching hash matching for video in April 2023’.

Following this initial response, Twitter subsequently provided eSafety with further information regarding tools to detect known CSEA video, and stated:

[Twitter] has continued to develop the technology required to support detection of known CSE video, that this is now in testing and [Twitter] anticipates being able to make this available imminently.

In response to a question about any alternative steps taken to detect known CSEA video, Twitter stated that it responds to ‘any reports received from users related to video content that contains CSAM material and immediately remove any reported content’ and that it ‘also conduct[s] proactive sweeps for content by skilled CSEA policy specialists.’

New or ‘unknown’ CSEA material

In response to questions about whether Twitter uses any tools to detect new CSEA material, Twitter stated that it does use tools on public posts but not on direct messages.

Table C

Service	Tools used to detect new CSEA?	Names of tools used
Tweets	Yes	Twitter did not provide the names of the tools in its original response, but subsequently confirmed it uses: Safer by Thorn Several internal tools including ProfileViewer2 and internal machine learning models and classifiers that do not have names.
Direct messages	No	

Twitter also responded that:

- Human moderators review all reports flagged by tools on public posts.
- An account that shares new CSEA material is permanently suspended and reported to NCMEC. All other accounts that engaged with that material are also suspended.

Twitter stated that tools are not in use to detect new CSEA material on direct messages because ‘the technology is still in development’, and that Twitter takes alternative steps such as using ‘a range of behavioural signals, in addition to content signals, to determine if accounts are violating our terms of service, as well as user reports’.

eSafety followed up with Twitter seeking clarification of Twitter's comment that technology is ‘still in development’, and whether Twitter was referring to different technology to that Twitter had stated it deploys on Tweets. Twitter responded that it is ‘still evaluating the privacy concerns and the efficacy of this technology to deploy on our service, given the novel challenges of identifying unhashed imagery’.

Following this response, Twitter subsequently provided eSafety with further information regarding measures to address new CSEA, and stated:

law enforcement agencies can also report to Twitter any potential CSEA violations that they receive regarding Direct Messages using Twitter’s legal request submission system.

URL blocking

Twitter stated that it blocks URLs linking to known CSEA material on the following parts of its service:

Table D

Parts of service	URLs linking to known CSEA material blocked?	URL sources
Public posts (‘tweets’)	Yes	NCMEC
Direct Messages	Yes	IWF Law enforcement Cybertip.ca Agents internal CSE investigations

Following a subsequent question from eSafety, providing Twitter with a further opportunity to provide the information sought in the Notice, Twitter added that in addition to NCMEC and IWF, it receives URLs from law enforcement branches from around the world and in Australia. Twitter added that it ‘continually explores other partnerships.’

Following this response, Twitter subsequently provided eSafety with further information regarding URL sources, and stated:

[Twitter] also receives URLs of known CSE from Canada’s Cybertip.ca... Twitter agents are also able to proactively block URLs in the context of internal CSE investigations.

Proportion of CSEA detected proactively

Twitter was asked what proportion of CSEA material is detected proactively, compared to CSEA material reported by users/trusted flaggers for public posts and direct messages. Twitter was asked to provide data for two periods of time within the Report Period, corresponding with the change of ownership in October 2022.

Table E

Time Period	Proportion of material detected proactively by Twitter (%)
24 January 2022 – 27 October 2022	90%
28 October 2022 – 31 January 2023	75%

Twitter noted that it was not able to provide data broken down by the part of the service (i.e. public posts and direct messages), so the numbers provided are for Twitter’s overall actions.

Twitter stated that it has improved its proactive detection methods since early 2022, which ‘helped us proactively detect more violative content as of November 2022 than is reported to us. We expect the proactive rate to improve further in February and March 2023’.

2. Questions about steps taken to identify whether an account that is found to have shared CSEA material includes a review of other material shared by that account

Twitter was asked whether it reviews an account for other CSEA material after it confirms that the account has shared CSEA material. In response, Twitter stated that it also reviews other material shared by that account – including a review of the public posts, and a review of direct messages of that account ‘to the extent it is necessary and proportionate’. Twitter stated that a combination of human moderator review and technical tools are used to identify CSEA content shared by violative accounts.

Following a subsequent question from eSafety, providing Twitter with a further opportunity to provide the information sought in the Notice, which required Twitter to name the tools used, Twitter responded that the tools include PhotoDNA, Thorn’s Safer and ‘other proprietary technologies’ including ProfileViewer2 as well as internal machine learning models and classifiers that do not have names.

3. Questions about measures in place to detect the livestreaming of child sexual exploitation and abuse on Twitter

Twitter was asked if it had any measures in place to detect the livestreaming of CSEA on its service. Twitter’s initial response was that Twitter no longer has livestream video following the discontinuation of the Periscope app.

Table F

Services	Measures in place to detect CSEA in livestreams or video calls/conferences	
	Yes/No	Names of tools used
Twitter	Twitter did not provide an answer regarding whether tools were used to detect CSEA livestreams during the Report Period. ⁹⁴	

eSafety followed-up with Twitter after their response to the Notice, providing Twitter with another opportunity to provide the information sought by the Notice. eSafety noted that Twitter was required to provide information on any measures in place during the Report Period. eSafety noted that as of 30 March 2023 Twitter retained a broadcast livestreaming feature. eSafety noted that Twitter's own account '@TwitterLive' was used to promote livestreamed broadcasts on the service. eSafety stated therefore that it did not consider it accurate for Twitter to report that it therefore 'no longer has livestream video following the discontinuation of the Periscope app'. eSafety also stated that its understanding was that other livestreaming features also existed on the Twitter service during the Report Period.

Twitter responded to eSafety's follow up question stating that, 'During the reporting period, Twitter deprecated the functionality for the public to be able to live stream content on Twitter, or on its legacy livestreaming product, Periscope.'

Twitter confirmed that 'Some pre-vetted media partners do have the ability to livestream events.'

In their subsequent response Twitter did not provide an answer to the Notice question outlining any measures it had to detect CSEA when livestreaming was available to all users during the Report Period, or any measures to detect CSEA when it was only available to 'pre-vetted media partners'.

Following this response, Twitter subsequently provided eSafety with further information regarding livestreaming on its service during the Report Period:

⁹⁴ Following publication of this Transparency Report X Corp. advised eSafety that its response to questions regarding the detection of livestreaming of CSEA on the Twitter service 'was inaccurate due to an inadvertent error' and provided a revised response to the question. X Corp.'s revised response is [available](#)

livestreaming was not available to all users through the reporting period. Noting the transition period to new ownership, Twitter deprecated the functionality and put this under review, limiting live video streaming to pre-vetted media partners only... if Twitter is made aware of or receives reports of potential CSE material happening via a video livestream, its teams review and respond, have the ability to suspend the livestream, permanently suspend the account and will report the account to NCMEC (Twitter has mandatory and dedicated NCMEC reporting avenues in place when there is an imminent risk to a minor). Twitter considers the fact that the livestreaming feature is available to a limited number of pre-vetted media partners only, in addition to the reporting mechanism, to be a mitigation measure that helps to prevent livestreaming of CSE in the platform.

4. Questions about steps to identify and remove underage users

Tools to detect underage users

In response to a question about what tools Twitter uses to detect underage users, Twitter stated that it requires all new accounts to enter their birthdate in order to use the service. Twitter also stated that users can report underage users via a form.

Number of underage users identified

In response to a question about the number of underage users Twitter has identified and removed during the Report Period, Twitter advised that it identified and removed the following number of underage users:

- Globally: 703,561 users
- In Australia: 7,618 users

Twitter provided a number regarding the proportion of underage users detected proactively, that it subsequently indicated was unrelated to the information sought. eSafety has therefore not published the original incorrect response.

Twitter stated that, if an underage user is reported by a user, Twitter reviews the account for potential signals or statements indicating the user is under 13, and 'if necessary' will request the user (or a parent/legal guardian) to provide a valid ID.

In response to a question about what measures Twitter has in place to ensure users whose accounts are closed for being underage are prevented from opening a new account, Twitter

stated that it proactively searches for accounts that may be linked to accounts that were previously suspended.

Following a subsequent question from eSafety, Twitter added that it uses an ‘internally built tool to identify accounts that share identifying pieces of data to automatically suspend accounts linked to previously suspended accounts.’

5. Questions about language analysis technology used to detect CSEA activity

Language analysis to detect grooming

Twitter was asked whether it uses any language analysis technology to detect likely online grooming on public posts or direct messages:

Table G

Service	Language analysis technology used to detect grooming?
Tweets	No
Direct messages	No

In response to a question about why no language analysis tools are used to detect grooming, Twitter responded that ‘Tweets are overwhelmingly short text messages, which are often not part of a conversation. We continue to monitor the development of technology in this area, for example by gaming services, but currently it is not of sufficient capability or accuracy to be deployed on Twitter’.

In response to a question about what alternative steps Twitter is taking to ensure grooming does not occur on its service, Twitter stated that ‘Twitter is not a service used by large numbers of young people, however we recognise that we need policies to protect against this’. Twitter stated that its child sexual exploitation policy also prohibits:

- Sending sexually explicit media to a child
- Engaging or trying to engage a child in a sexually explicit conversation
- Trying to obtain sexually explicit media from a child or trying to engage a child in sexual activity through blackmail or other incentives
- Promoting or normalising sexual attraction to minors as a form of identity or sexual orientation.

Twitter also stated that if an account is flagged for involvement in online grooming, Twitter permanently suspends that account and reports it to law enforcement.

Following this initial response, Twitter subsequently provided eSafety with further information regarding the steps it takes to ensure grooming does not occur on its services, and stated that Direct Messages ‘at sign-up default to allow messages only from people you follow, a mitigation measure which assists in preventing grooming from occurring’ on Twitter.

Language analysis technology to detect likely CSEA activity such as sexual extortion and the trading and sale of CSEA material

In response to a question about whether Twitter uses any language analysis technology to detect terms, abbreviations, codes and hashtags indicating likely CSEA activity, in particular but not limited to sexual extortion and the trading and sale of CSEA material, Twitter responded:

Table H

Service	Tools used to detect likely CSEA activity	Name of tools used
Public posts (‘tweets’)	Yes	‘Proprietary internal tools’
Direct messages	No	

When asked whether human moderators review all content flagged by these tools, Twitter responded that the majority of accounts that are flagged for CSEA are reviewed by a human moderator. Twitter added that it has a ‘small number of high precision signals’ where Twitter suspends users before a human reviews the content in order to quickly act on violating accounts. Twitter stated that, when indicators of CSEA activity are detected on an account, ‘either an account is human reviewed and then permanently suspended, or an account is automatically permanently suspended. In both cases we enqueue the account to also be sent on to law enforcement’.

Twitter was asked to provide information on why language analysis technology is not used for any part of the service (in this case, for ‘direct messages’ given Twitter’s answer above). Twitter was also asked to explain what alternative reasonable steps Twitter is taking to proactively minimise CSEA activity such as sexual extortion on its service.

Twitter did not provide an answer to either question.

Following a subsequent question from eSafety, providing Twitter with a further opportunity to provide the information sought in the Notice, Twitter stated that:

Twitter notes this technology is still new and in development, and decisions to deploy technology that will generate false positives must be balanced against the privacy intrusion of scanning private messages. We are monitoring and evaluating the applicability of developments for [Direct Messages] and remain committed to exploring new areas and options that will reliably solve these common challenges. Twitter remains open to privacy preserving options for our users that are right-sized for companies like ours and notes that reasonable training data is also needed for such tools to be effective, data which may not be available to smaller services or where the incidence of such activity is low. User reports, at this time, still best address the issues for further minimizing CSEA activity.

Following this response, Twitter subsequently provided eSafety with further information regarding the steps it takes when indicators of likely CSEA activity such as sexual extortion and the trading and sale of CSEA material are detected on an account. Twitter stated that, when it removes content, it ‘immediately reports it’ to NCMEC.

6. Question about information collected from users on sign up

Information collected at sign up

In response to a question about what user identifying information is collected on sign up to Twitter, Twitter responded that users are required to enter a phone number or email address. Twitter also stated that it collects background signals, such as IP address and interaction with other accounts.

Following a subsequent question from eSafety, providing Twitter with a further opportunity to provide the information sought in the Notice with regards to whether Twitter had provided all information collected at sign up as required by the question, Twitter added that it also collects the following information from users at sign up:

- A display name (for example, ‘Twitter’)
- A username (for example, @Twitter)
- A password
- Date of birth
- Display language

- Third-party single sign-in information (if the user chooses this sign-in method).

Twitter reported that it collects information about the Twitter settings that the individual selects, so they can respect those preferences, and stated that this is communicated in Twitter's privacy policy. Twitter added that if a user creates a professional account, they also need to provide Twitter with a professional category, and that additional details (such as payment information) are required to purchase ads or other offerings.

Authentication of accounts

Twitter was asked if it authenticates the information that it collects on sign up. Twitter responded that 'initially, users are required to complete three CAPTCHA challenges. Following this step, a verification email or SMS is sent to the email address or phone number registered to the account'.

Twitter also stated that it implements additional controls to help ensure the authenticity of users after sign-up, as follows:

- Twitter implements graduated access checks which may limit a user's reach while Twitter evaluates whether the account is operated authentically. When Twitter observes more account usages, the account will be unlocked to improve its accessibility and reach.
- Twitter Blue⁹⁵ subscribers need to verify themselves using their phone number and fulfil certain joining criteria. Historically, Twitter managed another verification process that required individuals to submit hard identifiers to verify their authenticity.

Following this response, Twitter subsequently provided eSafety with further information regarding authentication processes to ensure the account is controlled by the user signing up, and stated:

⁹⁵ Twitter Blue is a paid for subscription the provides accounts with certain benefits: [About Twitter Blue](#).

profile labels and checkmarks information continue to be updated. Twitter provides subscriptions such as Twitter Blue and also Verified Organizations⁹⁶ that require accounts to verify themselves. Additionally, prospective advertisers and eligible government accounts are provided complimentary checkmarks and are verified in this process.

7. Question about steps to prevent recidivism

In response to a question about what measures Twitter has in place to prevent recidivism of users for CSEA-related breaches on its services, Twitter responded that it uses proactive models and detection methods to detect ban evasion which consists of looking at specific data related to the original account. Twitter provided a link to its ban evasion policy.

Twitter identified multiple⁹⁷ indicators that it uses to prevent recidivism, which eSafety has chosen not to publish to prevent the information being misused.

Twitter was asked, if any information collected on sign up (outlined in section 5 above) was not used to detect recidivism, to provide reasons why not. Twitter did not answer the question with any reasons, but pointed to its privacy policy.

Following this response, Twitter subsequently provided eSafety with further information regarding information collected on sign up and prevention of recidivism. eSafety has decided not to publish to avoid this information being misused.

8. Questions about reporting in relation to CSEA material and activity

'In-service' reporting

Twitter was asked if users can report instances of CSEA activity or material to Twitter within the service (as opposed to navigating to a separate webform or email address). Twitter responded:

⁹⁶ Twitter Verification for organisations was available globally on 31 March 2023, so was not available during the Report Period of this Notice: [Twitter Verification for Organizations, a new kind of network on Twitter](#), 30 March 2023.

⁹⁷ eSafety uses the following terms to give an impression of the extent of the indicators used by services in the table below, rather than publishing the specific indicators which could be misused:

- Minimal: A small number
- Several: A moderate number
- Multiple: A significant number.

Table I

Service	In-service reporting option
Public posts ('tweets')	Yes
Direct messages	No

In relation to reporting content in direct messages, Twitter responded that 'users can still report DMs for CSEA through our web form'.⁹⁸

Median time for a user report of CSEA to be actioned

Twitter was asked to provide the median time taken to respond⁹⁹ to a user report about CSEA for the following parts of its service:

Table J

Provider	Services	Median time for user reported CSEA material to be actioned
Twitter	Public posts ('tweets') Direct messages	Twitter did not respond.

Following a subsequent question from eSafety, providing Twitter with a further opportunity to provide the information sought in the Notice, Twitter did not provide the median time. Twitter instead stated that it responds to 90% of child sexual exploitation user reports within 24 hours. Twitter added that in some cases, user reports require longer investigations and/or research which can delay a response.

9. Question about implementing end-to-end encryption on Twitter

Twitter was asked if a decision has been made to implement end-to-end encryption of direct messages during the Report Period. Twitter responded 'yes'.

Twitter was asked to detail any measures that were considered, or intended to be implemented, to ensure that end-to-end encryption does not increase the risk of CSEA dissemination without detection. Twitter responded 'Encrypted [direct messages] will not

⁹⁸ [Twitter Help Centre; Staying safe on X and sensitive content](#)
⁹⁹ Defined in the Notice as "Content removed and reported, user banned, or other content moderation decision made."

be available to all users at launch. Twitter will continue to leverage behavioral and other available signals, in addition to user reports, to take action on violations of our rules’.

eSafety notes that since the end of the Report Period Twitter implemented the end-to-end encryption of direct messages for some users.

Following this initial response, Twitter subsequently provided eSafety with further information regarding the measures considered, or intended to be implemented, to ensure that end-to-end encryption does not increase the risk of CSEA dissemination without detection, and stated:

users need to satisfy the following conditions in order to send and receive encrypted messages:

- both sender and recipient must be on the latest Twitter apps (iOS, Android, Web);
- both sender and recipient must be verified users or affiliated to a verified organization¹⁰⁰; and
- the recipient follows sender, or has sent a message to sender previously, or has accepted a Direct Message request from the sender before.

Twitter additionally clarified that, at this phase of launch, for now, encrypted messages can only be sent to one recipient, cannot include media such as images or links (i.e. can only be text) and have a range of other limitations at this time.

Twitter also restated that Direct Messages default at sign-up to allow messages only from people that a user follows.

10. Question about information provided in NCMEC reports

In response to a question about the information Twitter includes in its reports to NCMEC, Twitter provided a list of information that it includes in a report to NCMEC. eSafety has correlated this with the types of information NCMEC defines as ‘actionable’ for law enforcement.

¹⁰⁰ Twitter Verification for organisations was available globally on 31 March 2023, so was not available during the Report Period of this Notice. [Twitter Verification for Organizations, a new kind of network on Twitter, 30 March 2023](#)

11. Questions about Trust and Safety, and other staff at Twitter

Twitter was asked to provide the number of staff employed or contracted by Twitter to carry out certain functions between two specified periods – 24 January 2022 – 27 October 2022 and from 28 October 2022 – 31 January 2023, reflecting Twitter’s change of ownership.

Table K

Category of staff	Number of Staff Employed or Contracted by Twitter Between the Following Periods	
	Time period: 24 January 2022 - 27 October 2022	Time period: 28 October 2022 – 31 January 2023
Engineers focussed on trust and safety issues globally.	Twitter did not provide an answer.	Twitter did not provide an answer.
Trust and safety staff dedicated to CSEA issues globally.	Twitter did not provide an answer.	Twitter did not provide an answer.
Trust and Safety staff globally.	Twitter did not provide an answer	Twitter did not provide an answer.
Trust and safety staff in the APAC region.	Twitter did not provide an answer.	Twitter did not provide an answer.
Trust and safety staff in Australia.	Twitter did not provide an answer.	Twitter did not provide an answer.
Content moderators globally.	Twitter did not provide an answer.	Twitter did not provide an answer.
Content moderators in the Asia Pacific region.	Twitter did not provide an answer.	Twitter did not provide an answer.
Public policy staff globally.	Twitter did not provide an answer.	Twitter did not provide an answer.
Public policy staff in the APAC region.	Twitter did not provide an answer.	Twitter did not provide an answer.
Public policy staff in Australia.	2 (led by a regional team based in Singapore)	0

Twitter added that:

Twitter is engaged in a wide-ranging transformation and restructuring, to put the company on a path to financial stability. During this process, the company's total size has been significantly reduced, however the trust and safety function has been impacted to a much smaller degree than other functions. We currently have more than 3,000 people working on Trust and Safety at Twitter, including a mix of full time employees, contractors, and third-party services. Currently Twitter does not have content moderation or public policy staff located in Australia. Prior to the change in ownership, Twitter did have two public policy staff in Australia, led by a regional team based in Singapore. We continue to have a presence in the region, with matching time zones, as well as around the clock coverage. Our teams have a combination country context, product and policy knowledge and specialist capabilities.

Twitter did not provide any additional detail about its trust and safety staffing numbers, including where they are employed or what trust and safety functions they fulfil.

Following a subsequent question from eSafety, providing Twitter with a further opportunity to provide the information sought in the Notice for the other 9 categories of staff that the Notice required information on, Twitter responded that:

During the reporting period, Twitter underwent a major corporate restructuring, reducing its total global headcount by approximately 80% and undertaking a major review of roles, responsibilities and corporate functions. As such, the delineation set out does not easily map to Twitter's new corporate structure, where individuals perform a range of tasks and functions across numerous disciplines and subject areas.

Following this response, Twitter provided the following additional information:

irrespective of the corporate restructuring and associated organizational changes which have taken place - undertaken in order to best position the company for the future – it has at all times retained its zero-tolerance policy towards child sexual exploitation. Further, its associated processes and procedures have remained in place and have continued to be enforced across the transition period.

It added:

[Twitter] will continue to assess and, if necessary, modify the structure and composition of its teams, adding roles and headcount where deemed necessary, to ensure its business is equipped with the right level of resources.

The company has staff dedicated to online trust and safety, individuals with knowledge, experience and context of Twitter, its service and the topics and organizations related to online safety, and Australia, whose responsibility it is to enforce its policies in this area.... Twitter's trust and safety experts have not been based in / situated on-ground in Australia, being instead located outside Australia, across Twitter office locations and in timezones around the world that overlap with all Australian timezones.

Twitter's expertise includes Australian and foreign nationals who have the ability to travel to liaise in person, as well as online, with e-Safety, other government and law enforcement agencies, and other organizations.

Twitter added that 'during this period of transition', 'Twitter has adjusted levels of engagement, but maintained its levels of commitment to ensuring Australians are safe online, and this remains the case.'

Twitter added that it 'continues to have employees based in Australia, who are mostly focused on Sales and Marketing duties.'

12. Question about the languages human moderators operate across

eSafety asked Twitter to list all the languages that Twitter's human moderators (employees and contractors) operate across. Twitter listed the following languages:

English; Spanish; Japanese; Turkish; Arabic; Hindi; Indonesian; French; Korean; German; Chinese; Portuguese

Twitter stated that language translation software is used to support remaining languages not covered by native speaking agents.

Following this initial response, Twitter subsequently provided eSafety with further information regarding the languages human moderators operate across, and stated:

[Twitter] continues to evolve its language capabilities as the service grows. The company has the ability to seek and conduct vendor services in a range of additional languages, which include but are not limited to those needed in the event of additional reviews, or for emergencies.

13. Questions about Twitter Recommender systems

Design objectives

In response to a question about the top design objectives of Twitter's recommender system, Twitter stated that 'unregretted user minutes' is Twitter's 'overall metric'. Twitter also provided a link to more information on its 'For you' feed.¹⁰¹

Following a subsequent question from eSafety, providing Twitter with a further opportunity to provide the information sought in the Notice, Twitter submitted that 'the top objectives for Twitter's Recommender Systems are to increase time spent on the platform and recommend content with:

- A high likelihood of engagement (likes, replies, bookmarks)
- Low likelihood of being reports of bad content
- Low likelihood of author being blocked'.

Signals used to recommend content

When asked to provide a list and description of the top signals that Twitter uses in determining what content should be recommended to users, and the relative importance of each signal in meeting the design objectives, Twitter responded 'Twitter will open source all code used to recommend tweets in the coming weeks.'

eSafety noted in a follow-up question, that a commitment to the future publication of partial code was not what the Notice had required and provided Twitter with a further opportunity to provide the information sought in the Notice. Twitter clarified that 'there are 9 relevant components that make up the 'for you' timeline, which have now been published.' Twitter stated the types of signals leveraged include:

- Number of likes
- Number of retweets

¹⁰¹ A recommender system – <https://help.twitter.com/en/using-twitter/twitter-timeline>.

- Views
- Number of replies
- Whether a user engaged with the author in the past
- Users' interests
- Recency of a tweet
- User's sensitive media settings (minors cannot view sensitive no matter their settings)
- Suspected authenticity ('not spammy, twitter blue').

Twitter noted that the nine relevant components that make up the 'for you' timeline had been [published](#). Twitter noted that it would continue to provide information and update the publication of the recommendation algorithm on Github.

Questions about risk impact assessments and auditing

Twitter was asked questions about the measures it employs to test and update its recommender system to improve the safety of its service and avoid amplifying harmful communities.

Twitter did not respond.

Following a subsequent question from eSafety, providing Twitter with a further opportunity to provide the information sought in the Notice, Twitter stated that it 'strives to ensure that its products are as safe as possible and conducts assessments and tests to ensure its recommendation algorithm is.' Twitter also stated 'our terms of service identify content and behaviours that are unacceptable, and we apply remediations when content violates our terms of service. We use internal metrics to determine the impact of each launch.'

Age-appropriate restrictions to recommendations

Twitter was asked what criteria it has for evaluating content that is harmful to children but that does not violate community guidelines for the purpose of not recommending or recommending the content less often. Twitter responded: 'Twitter's service is overwhelmingly not used by children. Taking action on those who seek to engage in child abuse is our number one priority'.

Twitter also stated that it continues to put in place a range of settings to limit amplification of, and reduce the risk of stumbling across, harmful content for those aged 13 – 18, including through default sensitive media settings for this cohort, and rules preventing

advertisers from showing inappropriate advertising to them (such as alcohol). Twitter noted that ‘The challenges here are significant; definitions and specific determinations of harmful content continue to carry significant concerns including real risks of limiting reasonable access to information, ideas and speech’.

Twitter did not respond to a question requiring the criteria used for evaluating content that is harmful to children.

Following a subsequent question from eSafety, providing Twitter with a further opportunity to provide the information sought in the Notice, Twitter noted:

Aside from evaluating content as [sensitive/graphic](#) or not, we do not currently perform any additional evaluation of content as being harmful for children, given that children are not our target customer and our service is not overwhelmingly used by children. Our priority is enforcing our rules, and providing appropriate safeguards and controls to users. Some content is gated behind an age appropriate content interstitial, based on whether a person has submitted an age under 18.

Our [help centre](#) states: ‘Age Restricted Content: We restrict viewers who are under 18, or who do not include a birth date on their profile, from viewing adult content. You can learn how to [add a birth date](#) to your profile, [adjust birth date visibility settings](#) (visibility of your birth date is defaulted to private if you update it after January 2022) and learn how Twitter uses your age to show you more relevant content, including ads, in accordance with our Privacy Policy. People over 18 can opt-out of viewing sensitive media on Twitter by updating their settings.’

Following this initial response, Twitter subsequently provided eSafety with further information and stated it:

continues to explore solutions that address age verification whilst at the same time balancing data processing minimisation principles and data protection concerns generally speaking, especially for minors.

Preventing the sale and trade of CSEA material

Twitter was asked what steps are taken to prevent the recommendation of accounts or content involved in the sale and trade of child sexual exploitation material. Twitter responded:

We do not allow CSAM material on Twitter and when we are aware of it we immediately remove it. When we identify content we also suspend accounts that have engaged with the materials. We also make known cse keywords unsearchable to avoid returning cse materials through search.

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

Twitter responded that it does have internal metrics in place to assess the efficacy of its interventions to detect CSEA material. Twitter stated that it measures what percent of accounts are taken down proactively versus by external partners (such as law enforcement or user reports), and that it tracks impressions before takedown. Twitter stated that it continues to monitor the precision of its models, as it stated that high-precision models allow Twitter to automate suspension without human review.

15. Question about appeals against CSEA-related moderation

In response to a question about how many appeals have been made by users for accounts banned or content removed for CSEA material and activity, Twitter responded ‘We see a consistent appeals rate of between 2-4% on all CSE related suspensions.’ Twitter stated that ‘a very low proportion (<10%) of those appeals are successful’.

eSafety provided Twitter with a further opportunity to provide the information sought in the Notice, noting that it was the number of appeals which was required by the Notice.

Twitter responded that between January 2022 and December 2022, it suspended 2,348,712 accounts for violating Twitter’s CSE policy, globally. Twitter added that in the same period, Twitter received via its internal redress mechanism 446,684 appeals associated with 238,352 unique Twitter accounts.

16. Questions about Safety by Design training for staff

Twitter responded that it does provide training to its staff on Safety by Design principles. Twitter stated, ‘For new features, trust and safety staff members work with teams across the company to proactively surface any risks and build in mitigating features.’

8.3 TikTok Summary

Overview

TikTok Pte. Ltd. was asked about its TikTok service.

Note: eSafety used the term child sexual exploitation and abuse (CSEA) in the Notice given to TikTok. However, TikTok's responses sometimes used the term CSAM (child sexual abuse material), rather than CSEA.

1. Questions about known and new CSEA material

Known CSEA images

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

Table A

Part of service	Hash matching to detect known images?	Names of tools
Profile pictures	Yes	PhotoDNA
Public photos	Yes	An internal TikTok model which 'consists of multiple TikTok machine learning algorithms that review videos, photo/images, audios and text-based content, and is conceptually equivalent to hash matching'.
Direct messages ¹⁰²	Yes	<p>TikTok stated in its initial response that users cannot share photos via direct message in Australia.</p> <p>Following a subsequent question from eSafety, seeking clarification of TikTok's response to the Notice, TikTok stated that users in Australia can only share images (via direct message) where the image has already been posted on the TikTok platform and where the post has not been set to private. TikTok stated that this content has already</p>

¹⁰² Relating to the sharing (via direct messages) of images that have already been posted to the TikTok platform and the settings of the post are not set to private. TikTok clarified that users are not able to upload images (or videos) to direct messages directly from their camera.

Part of service	Hash matching to detect known images?	Names of tools
		been through TikTok's moderation process (including hash matching) before it is shared. TikTok stated that users cannot share new images from their own device (e.g. camera roll) that have not been posted to the platform via direct message.
Private/Only Me photos	No	

TikTok reported that it sources its hashes of CSEA images from the following databases:

- Internet Watch Foundation (IWF) Library
- National Centre for Missing & Exploited Children (NCMEC) Library (NCMEC NGO List and NCMEC Industry List)
- Internal TikTok Library – a database of prior violations (including CSEA) which the Internal TikTok Model uses to match against content uploaded to the platform. TikTok also stated that through this database it employs ‘techniques based on the same principles as irreversible hash technology to ensure CSAM is not retained or stored.’

TikTok stated that it takes a subset of what is made available from these lists. TikTok explained that the hashes need to be in the format that matches the hashing techniques deployed by TikTok. TikTok reported that it takes a full set of PhotoDNA hashes from the relevant sources listed.

TikTok reported that it updates the hashes from these databases daily, ‘to ensure that we are ingesting the most up to date identified content.’

In response to why hash matching tools are not used on Private/only me photos and what alternative steps TikTok is taking to detect known CSEA images on this part of its service, TikTok stated that all content, including photos posted in private, that are uploaded to the platform are reviewed by ‘automated detection technology’. TikTok confirmed, in a clarifying question, that these tools do not involve hash matching. However, TikTok stated that it has approved the process of using hash matching tools over private photos and is currently implementing the tool:

If a potential violation is found, the automated moderation system will either remove it from the platform automatically, or where certain signals are present or a user reports it, it is sent to our human moderation teams for further review. Additionally, all CSEA violations are sent to a specialised team for confirmation and reporting to NCMEC.

Following this initial response, TikTok subsequently provided eSafety with further information regarding hash matching tools used on Private/only me photos, and stated:

TikTok has implemented hash matching tools over photos posted in the ‘Private’/‘Only Me’ setting and is currently troubleshooting its use in this part of the service.

Known CSEA video

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

Table B

Part of service	Hash matching to detect known video?	Name of tools used
Public videos Videos in direct messages ¹⁰³	Yes	PhotoDNA Google Content Safety API CSAI match Internal TikTok Model ¹⁰⁴
Private/Only Me videos	No	

TikTok stated that it has approved the use of hash matching tools over Private/Only Me videos and was in the process of implementing it on the service. In response to a question about why hash matching tools are not used on private/only me videos and what alternative steps TikTok is taking to detect known CSEA videos, it stated that the same steps are taken for known videos as those for Private/only me images.

¹⁰³ Relating to the sharing (via direct messages) of videos that have already been posted to the TikTok platform and the video was not posted in the ‘Private’/‘Only Me’ setting. TikTok clarified that as with images, users are not able to upload videos directly from their camera roll.

¹⁰⁴ TikTok reported that this includes machine learning algorithms that review videos, images, audio and text-based content and stated it was conceptually equivalent to hash matching.

Following this initial response, TikTok subsequently provided eSafety with further information regarding hash matching tools used on private/only me videos, and stated:

TikTok has implemented hash matching tools over videos posted in the ‘Private’/‘Only Me’ setting and is currently troubleshooting its use in this part of the service.

New or ‘unknown’ CSEA material

In response to questions about whether TikTok uses any tools to detect new CSEA material, TikTok responded that it uses the following tools:

Table C

Service	Tools used to detect new CSEA?	Names of tools used
Profile pictures	Yes	TikTok stated that it uses a combination of: Internal computer vision model ¹⁰⁵ Audio and NLP (Natural Language Progressing) model Google Content Safety API
Public photos and videos	Yes	
Private/Only Me photos and videos	Yes	
Photos and videos in direct messages ¹⁰⁶	Yes	
TikTok live video	Yes	

In response to a question about what action TikTok takes when an account is detected sharing or storing new CSEA material, TikTok outlined several steps it takes, including:

- Content is prevented from being uploaded to the platform or the post is removed from the platform;
- The account is permanently banned;
- Account and violating content are reported to NCMEC;

¹⁰⁵ TikTok defined this as a model that divides videos into frames, which are then processed (for example, to detect exposed body skin).
¹⁰⁶ Content that has already been uploaded to the platform and are sent via direct message.

- High risk cases trigger TikTok’s use of NCMEC’s industry escalation system and may result in direct outreach to law enforcement for immediate intervention and coordination;
- Other steps that eSafety has decided not to publish to avoid the information being misused.

TikTok reported that human moderators review all new CSEA content that is flagged by tools on the services listed above.

URL blocking

In answer to a question about whether TikTok blocks URLs linking to known CSEA material, TikTok provided the following information.

Table D

Parts of service	URLs linking to known CSEA material blocked?	URL sources
Links overlayed in photos/videos	Yes	IWF
Links in photo/video descriptions	Yes	Internal TikTok Library ¹⁰⁷
Links sent in direct messages	No	

TikTok stated that it sources its URLs linking to known CSEA from the IWF on a daily basis and from its internal TikTok Library. It noted the following about the TikTok Library:

we maintain an internal list of URL links to suspected CSAM, which [TikTok] proactively block[s] and remove[s] from links overlayed in photos/videos, or in photo/video descriptions. [TikTok] also report[s] these to NCMEC for further investigation.

In response to why URLs linking to known CSEA material are not blocked on links sent in direct messages TikTok explained the live URL hyperlink is disabled when sent via direct message so that users cannot use the URL to click through to the webpage. eSafety notes that a URL can however be copied from a direct message into a browser and followed.

¹⁰⁷ Captured from prior violations.

TikTok stated that it is working on a product feature to prevent URLs to known CSEA material being sent via direct messages and anticipated that the development of this feature would be completed in 2023.

Proportion of CSEA detected proactively

TikTok was asked what proportion of CSEA material is detected proactively, compared to CSEA material reported by users/trusted flaggers. TikTok provided statistics for part of the Report Period.

Following a subsequent question from eSafety, seeking clarification of TikTok's response to the Notice, TikTok responded that 'the proactive detection rate for CSEA material detected on the platform during the Report Period is approximately 94.4%. The figures are global, however, exclude data from the United States.'

TikTok also provided broader data, which it said became available after its original notice response. TikTok explained that while 'Minor Safety sub-policies capture CSEA material, they also reflect Minor Safety violations that are unrelated to CSEA (due to the nature of each policy title).'

Table E

		Jan-Mar 2022	Apr-Jun 2022	Jul-Sept 2022	Oct-Dec 2022
Nudity and sexual activity	Proactively detected	96.8%	96.9%	97.0%	96.7%
	Incurred a zero view rate	92.6%	91.4%	87.2%	79.7%
Sexual exploitation of minors	Proactively detected	90.6%	93.2%	95.1%	93.1%
	Incurred a zero view rate	82.5%	85.8%	88.3%	82.1%
Grooming behaviour	Proactively detected	96.6%	96.7%	96.6%	96.2%
	Incurred a zero view rate	90.6%	88.4%	83.4%	73.9%

2. Questions about steps taken to identify whether an account that has been found to have shared CSEA material contains other CSEA material

TikTok was asked whether it reviews an account for other CSEA material after it confirms that the account has shared CSEA material. TikTok stated that once it confirms an account has shared CSEA material, the account is reported to NCMEC and the data is held to allow for law enforcement disclosure, in line with TikTok's law Enforcement Guidelines. TikTok noted that 'if there is reason to believe additional violations may be found, a subset of the account may also be further reviewed by [TikTok's] specialist teams.' TikTok stated this review may include:

- Public and private video and image posts
- Livestreams
- Comments posted by the account holder
- Account holder's 'profile' (which includes profile picture/avatar, name and bio)
- Direct messages (where there are compelling grounds, such as imminent risk to life/safety, access to direct message may be sought, which is subject to legal approval under applicable laws). TikTok stated that this action is not undertaken in all jurisdictions.

TikTok also stated that human moderators look into other parts of the account and that it generally does not use technical tools for this process. TikTok stated that 'For the purpose of looking into other content/parts of the account, after an account is confirmed as having shared CSEA material, we do not generally use tools because for this process we rely on our specialist human moderation teams.'

3. Questions about measures in place to detect the livestreaming of child sexual exploitation and abuse on TikTok

TikTok was asked if it had any measures in place to detect the livestreaming of CSEA on its service. TikTok responded that it does have measures in place and uses the following tools:

Table F

Services	Measures in place to detect CSEA in livestreams or video calls/conferences	
	Yes/No	Names of tools used
TikTok live video	Yes	Internal computer vision model Internal audio and NLP model Machine learning behavioural model

TikTok added that the internal computer vision model is a model that divides videos into frames, which are then processed (for example, to detect exposed body skin).

TikTok stated that human moderators review all CSEA reports flagged by these tools, as well as reports from users.

TikTok also listed a variety of indicators that it uses to detect livestreaming of CSEA, which eSafety has chosen not to publish to avoid the information being misused.

In answer to a question about what steps are taken when an account is detected for livestreaming CSEA, TikTok stated that the livestream is stopped and is unable to be viewed by users, the host and participants engaging in the livestream (through sending gifts, commenting or liking) are permanently banned, their accounts and the livestream content is sent to NCMEC, and high risk cases may be sent directly to law enforcement for immediate action via NCMEC's ESP escalation system. Other steps are taken by TikTok that eSafety has decided not to publish to avoid the information being misused.

4. Questions about steps to identify and remove underage users

Tools to detect underage users

TikTok stated that it considers those under the age of 13 as underage users. When asked what tools are used to detect underage users, TikTok responded that users are required to self-declare their age at the time of registration through inputting their date of birth. TikTok also reported that it uses proactive detection through technology and human moderation to 'detect and remove underage users from the platform' which includes language analysis that is then reviewed by human moderators. Third parties and users are also able to report suspected underage accounts through the 'report account' feature.

Number of underage users identified

In response to a question about the number of underage users TikTok has identified and removed during the Report Period, TikTok advised that it identified and removed the following number of underage users:

Table G

Region	Number of underage users removed	Proportion detected proactively ¹⁰⁸
Globally	79,312,821 users	80.6%
Australia	963,158 users	86.7%

In answer to a question about the methods used by TikTok to determine if a user report of an underage account was indeed created by an underage user, TikTok responded that its internal specialist teams assess the account and its content against the “Underage Banning Guidelines”.

5. Questions about language analysis technology used to detect CSEA activity

Language analysis technology to detect likely online grooming

TikTok was asked whether it uses language analysis technology to detect likely grooming. TikTok reported that it uses automated tools on comments on photos/videos, direct messages and comments on TikTok Live videos. TikTok stated that it uses the Internal TikTok NLP model and a language analysis model in order to perform its detection capability.

TikTok reported (including following subsequent questions from eSafety) that the approach for whether human moderators review all reports flagged by technology is different depending on the part of the service and the degree of certainty and types of risk identified by that technology.

¹⁰⁸ TikTok stated that all other underage user removals that were not detected proactively were a result of user or third-party reporting.

Table H

Part of TikTok Service	Human moderators review of all, or a portion of all reports, flagged by technology
Comments on photos, videos and livestreams	<p>TikTok stated that if the technology flags the comment as high risk and there is a high degree of certainty that it violated Community Guidelines, the comment is automatically removed, and human moderators do not review.</p> <p>If the technology flags the comments as being potentially violative of Community Guidelines and the certainty predicted is not high, the comment is sent for human moderator review.</p> <p>All comments that are reported by a user as suspected online grooming are reviewed by human moderators.</p> <p>In response to a follow-up question about the proportion of reports viewed by human moderators, TikTok stated that almost 100% of comments that are flagged by technology as potentially violative of policies including online grooming and likely CSEA activity are actioned by technology and that less than 1% are reviewed by human moderators.</p>
Direct messages	<p>TikTok stated that depending on the nature of the violation, the technology will either:</p> <ul style="list-style-type: none"> Apply a message ban, where the sender will receive a notification in the message conversation to inform them that the message is not sent, and the recipient will not receive the message. In this case, human moderation is only performed if a user appeals the message ban. Send the recipient a pop-up noting that the message may be inappropriate and ask the user whether they want to report the message. If they do make a report, it is reviewed by human moderators. <p>In response to a follow up question from eSafety seeking clarification on TikTok's response, TikTok responded that 0% of direct messages flagged by technology are sent for review by human moderators.</p> <p>TikTok stated that to reach a balance between privacy obligations and user safety, technology detection automatically interrupts the flow of communication between users and 'are designed to encourage a pathway that will lead to the material being reviewed by human moderators.' TikTok stated that all reports made by a user are reviewed by human moderators.</p>

Language analysis technology to detect likely terms, abbreviations, codes and hashtags indicating likely CSEA activity

TikTok was asked whether it uses language analysis technology to detect likely terms, abbreviations, codes and hashtags indicating likely CSEA activity, in particular but not

limited to, the exchange of gifts for underage sexual acts and sexual comments, and sexual extortion.

TikTok stated tools are used on photos/videos, direct messages and comments on TikTok Live videos, using its Internal TikTok NLP tool (Natural Language Processing).

TikTok added that it sources terms and abbreviations from the IWF's Library, and from NGOs (Thai government agency, IJM (International Justice Mission) Philippines), TikTok's in-house intelligence courses (Trust & Safety, Risk and Analysis teams) and the Internal TikTok Library.

Regarding whether human moderators review all or a portion of all reports flagged by the technology, TikTok stated (including following subsequent questions from eSafety) that it follows the same process detailed above for its human moderator review for detection of likely online grooming.

In response to what steps it takes when indicators of likely CSEA activity are detected on an account, TikTok reported that it differentiates between services, as follows:

Table I

Part of TikTok service	Steps taken when indicators of CSEA are detected on an account
Comments on photos, videos and livestreams	<ul style="list-style-type: none"> Automatically removed from the platform by technology; or Queued for human review by specialist teams who review against TikTok's Minor Safety policies (which include CSEA activity). If CSEA activity is identified, the same steps are taken as when account is detected sharing or storing new CSEA material and when an account is detected for livestreaming CSEA as detailed above. This includes reporting to NCMEC or escalated to NCMEC's ESP system or outreach to law enforcement.
Direct messages	<ul style="list-style-type: none"> Block the message from being sent; or Prompt the recipient to report the message which, if done, is reviewed by human moderators. Send the recipient a pop-up noting that the message may be inappropriate and ask the user whether they want to report the message. If they do make a report, it is reviewed by human moderators. If CSEA activity is identified, the same steps are taken as when account is detected sharing or storing new CSEA material and when an account is detected for livestreaming CSEA as detailed above. This includes reporting to NCMEC or escalated to NCMEC's ESP system or outreach to law enforcement.

6. Question about information collected from users on sign up

In response to what identification information is collected from users on sign up, TikTok listed the following which is also outlined in its Privacy Policy:

- Date of birth
- Username
- Information about the user's network or device (e.g. IP address or approximate location)
- Phone number or email address or, if signing up using a third party platform (e.g. Facebook, Google, Twitter, Apple or Instagram), account information such as their username and public profile or other possible information related to such accounts
- Name associated with the account, profile photo or bio.

Authentication of accounts

TikTok was asked if any of the above information collected at sign up was authenticated. TikTok stated that the process of account authentication at the time of sign up differs based on the method of sign up (for example if a user signs up by email, an email is sent asking the user to verify it),

TikTok stated that users can also sign up using an existing social media account, which then utilises the single sign on method via Facebook, Google, Twitter, Apple or Instagram.

7. Questions about steps to prevent recidivism

In response to a question about what measures TikTok has in place to prevent recidivism of users banned for CSEA-related breaches on its services, TikTok identified multiple¹⁰⁹ indicators that it uses to prevent recidivism. eSafety has chosen not to publish these indicators.

TikTok was asked, if any information collected on sign up (outlined in section 6 above) was not used to detect recidivism, to provide reasons why not. TikTok explained why some

¹⁰⁹ eSafety uses the following terms to give an impression of the extent of the indicators used by services in the table below, rather than publishing the specific indicators which could be misused:

- Minimal: A small number
- Several: A moderate number
- Multiple: A significant number.

information was not used, which eSafety has decided not to publish to avoid the information being misused.

8. Questions about reporting in relation to CSEA material and activity

‘In-service’ reporting

TikTok was asked if users can report instances of CSEA activity or material to TikTok within the service (as opposed to navigating to a separate webform or email address). TikTok responded that users can report photos/videos shared publicly (to everyone or followers only), TikTok Live and direct messages in-service.

Median time for a user report of CSEA to be actioned

TikTok was asked to provide the median time taken to respond¹¹⁰ to a user report about CSEA for the following parts of its service:

Table J

Part of service	Median time for user reports CSEA material to be actioned
Photos/videos share publicly (to everyone or followers only)	5.2 minutes
TikTok Live	7.7 minutes
Direct messages	7.4 hours

9. Question about what languages human moderators operate across

TikTok was asked to list all the languages that its moderators operate across. In its initial response to the question of what languages does TikTok’s human moderators operate across, TikTok responded over 70 languages, but did not specify which languages. Following a subsequent question from eSafety, providing TikTok with a further opportunity to provide a list of the languages (rather than the total number), TikTok provided the following list:

¹¹⁰ Defined in the Notice as “Content removed and reported, user banned, or other content moderation decision made.”

Table K

Afrikaans	Filipino	Kazakh	Russian	Zulu
Amharic	Finnish	Khmer	Serbian	
Arabic	French	Korean	Sesotho	
Assamese	Garhwali	Kurdish	Sinhalese	
Azerbaijani	German	Lao	Somali	
Bengali	Greek	Malay	Spanish	
Bhojpuri	Gujarati	Malayalam	Swahili	
Brazilian Portuguese	Haryanvi	Mandarin	Swedish	
Bulgarian	Hausa	Marathi	Tamil	
Burmese	Hebrew	Minnan	Telugu	
Cantonese	Hindi	Nepali	Thai	
Croatian	Hungarian	Norwegian	Turkish	
Czech	Icelandic	Odia	Twi	
Danish	Igbo	Persian	Ukrainian	
Dutch	Indonesian	Polish	Urdu	
English	Italian	Punjabi	Vietnamese	
Estonian	Japanese	Rajasthani	Xhosa	
European Portuguese	Kannada	Romanian	Yoruba	

10. Questions about the exploitation of TikTok's 'Only Me' content visibility setting

eSafety asked if TikTok uses any tools, systems, or signals to monitor account credential activity and flag suspected account sharing for review by moderators. eSafety noted that a 2022 Forbes report highlighted that TikTok's 'Only Me' content-visibility setting was being exploited to create 'closed environments' where CSEA material could be shared between people who accessed private 'Only Me' posts using shared account credentials. eSafety also

noted that TikTok guidelines prohibit users from providing ‘access to your account credentials to others’.

TikTok responded by stating that it does use tools, systems and signals to monitor account credential activity and flags suspected account sharing for review by moderators. TikTok stated that it has policies in place to prohibit users from providing others with access to account credentials as this activity is against Community Guidelines. TikTok stated that it implements detection tools to ensure common signals are detected. TikTok responded that ‘Content and associated users that are identified as being violative of these policies will have enforcement action taken at an account level to ensure that the user is not able to produce similar content by creating new accounts.’ TikTok reported it also takes other steps that eSafety has decided not to publish to avoid the information being misused.

Number of accounts TikTok removed for sharing account credentials

TikTok reported that while the sharing of account credentials violates its Community Guidelines, TikTok does not typically track data in relation to the violations eSafety specified. Instead, TikTok provided the number of account bans in Australia during the notice Report Period.

- Account bans: 4,805

TikTok provided the following details for the figure above:

- Includes detection and moderation of the ‘post to private’ trend by TikTok specialist teams (who perform sweeps across the platform) and does not include the violations detected and actioned by TikTok proactive detection models;
- May include some data associated with other violations (i.e. non ‘post to private’) because it was not possible to breakdown all violation data, due to the way in which TikTok track, compile and store data; and
- Reflects accounts that were banned for sharing account credentials and accounts that signalled participation or attempts to solicit participation in ‘post to private’ (e.g. by sharing a relevant hashtag).

In addition, eSafety asked TikTok to provide the number of accounts removed for sharing account credentials who were also found to be sharing CSEA material during the Report Period. TikTok stated that the data requested was not available for the full Report Period due to its data retention practices but provided the following figure for 1 August 2022 to 31 January 2023:

- Accounts also sharing CSEA material: 37

Tik Tok also noted that the figures include CSEA-related violations, not only violations for sharing CSEA material.

Implementation of measures to detect TikTok accounts that publicly encourage other users to request account credentials

eSafety asked if TikTok had implemented any new measures, or if it had updated any existing measures, to detect TikTok accounts that publicly encourage other users to request account credentials in order to access content stored privately using the ‘Only Me’ setting – a.k.a. ‘posting-in-private’ accounts.

TikTok stated that prior to the issue of sharing of account credentials being raised, its Community Guidelines already prohibited users from providing others with account access credentials. TikTok also noted that on 21 March 2023, it updated the Community Guidelines, and these changes came into effect on 21 April 2023. TikTok reported that to address the issue of users encouraging others to request account credentials, it has improved its detection strategy of these violations through technology moderation. It also explained that it blocks terms based on language analysis and has issued specific guidance and training to moderation teams so that they are alerted to these signals. ‘TikTok has also updated the reporting interface for users, so that the “Minor Safety” category appears at the top of the list when users select a reason for reporting a video, livestream, account or comment.’ TikTok noted these measures are in addition to the tools and systems it already has in place to detect ‘CSAM’.

11. Question about implementing end-to-end encryption on TikTok

TikTok stated that it has not carried out work with the goal of implementing end-to-end encryption on any part of its service during the Report Period.

12. Question about information provided in NCMEC reports

In response to a question about the information TikTok includes in its reports to NCMEC, TikTok provided a list of information that it includes in a report to NCMEC. eSafety has correlated this with the types of information NCMEC defines as ‘actionable’ for law enforcement.

13. Questions about TikTok recommender systems

Design objectives of TikTok’s algorithmic recommender systems (e.g. “For You”).

TikTok stated that the objectives of its recommender system for the For You Feed (‘FYF’) is to provide content that is personalised and of interest to users, provide diverse content, and ensure safety and quality is maintained for all users. TikTok also reported that it prioritises inspiring creativity and bringing joy and the FYF plays a large role in this. The FYF is empowered by multiple machine learning models and rule engines to deliver content that is likely to be of interest to the user.

TikTok added that its ‘recommendation system is also designed with safety as a consideration. Because the FYF is intended for a general audience, certain content, while allowable under the Community Guidelines, is not eligible for recommendation. In [TikTok’s] latest Community Guidelines to be effective on April 21, 2023’, TikTok identified that it makes ineligible for the FYF certain content that is related to: (1) Behavioural Health, (2) Sensitive and Mature Themes, (3) Integrity and Authenticity, and (4) Regulated Goods. Further details can be found in the Community Guidelines and in response to [another question of the notice pertaining to evaluation of recommended content].’

Signals used to recommend content

When asked to provide a list and description of the top signals that TikTok uses in determining what content should be recommended to users, and the relative importance of each signal in meeting the design objectives, TikTok responded that its machine learning models predict a user’s potential engagement with a video using signals including:

- User engagement (video playtime, likes, shares, accounts followed, comments and content created)
- Account and device information (language preference, country settings, device type)
- Video information (captions, sounds, hashtags)

TikTok stated that ‘a strong indicator of interest, such as whether a user finishes watching a longer video from beginning to end, would receive greater weight than a weak indicator, such as whether the video's viewer and creator are both in the same country. These weights change dynamically as the models “learn” more about a particular user’s behaviour.’ TikTok also stated that it has publicly released information to users regarding why a particular video was recommended to them ([Learn why a video is recommended For You | TikTok Newsroom](#)).

14. Questions about risk impact assessments and auditing

TikTok was asked questions about the measures it employs to test and update its system to improve the safety of its service and avoid amplifying harmful communities. TikTok responded that its Trust and Safety teams use a range of measures to maintain and improve the overall safety of the platform, including through content moderation and quality assurance; training of its moderation teams; user report and user appeals; specialist moderation queues; collaboration with experts across industry; and specialised investigations into incidents.

In answer to a question about whether tests and updates are performed and how often they are performed, TikTok reported that:

Tests and updates are performed regularly on a daily business as usual (“BAU”) basis, including content moderation coverage 24 hours per day, 7 days a week. In addition to onboarding training for moderators, training is conducted regularly and on as needed basis to ensure the moderators keep pace with dynamic online content, latest trends and current issues.

15. Question about age-appropriate restrictions to recommendations

TikTok was asked what criteria it has for evaluating content that is harmful to children but that does not violate community guidelines (e.g. content that contains implied nudity, sexualises body parts or is blatantly erotic or sensual) for the purpose of not recommending or recommending the content less often. TikTok responded that the measures it has in place to test and update the TikTok system (i.e. content moderation and quality assurance; training; user report & user appeal; specialist moderation queues; collaboration with experts; and specialised investigations into incidents) form the criteria for evaluating age-appropriate content restrictions to recommendations.

TikTok also added that it commenced a phased implementation roll out of its content classification system in 2022. TikTok noted that this involved sorting video content into content categories based on maturity themes. TikTok added that 'Content rated as being generally suitable for audiences aged 18+ (but that does not violate the Community Guidelines), is restricted to 18+ users, to help prevent certain content from reaching audiences between the ages of 13-17, either via recommendations on the FYF, via search or through content being shared with that user. If a user aged under 18 years of age attempts to access this type of content, a screen will appear that states, 'Post unavailable' and 'This post is age-protected'.

In addition, TikTok stated that non-exhaustive criteria in its Community Guidelines during the notice Report Period and its subsequently updated Community Guidelines are used to assess whether content may be harmful to children and should not be eligible for recommendation, including:

- Minor Safety: content uploaded by users under the age of 16 is not eligible for recommendation
- Overtly sexualized or sexually suggestive content: Content that is overtly sexually suggestive may not be eligible for recommendation. This would include content that depicts implied nudity, sexualizes body parts, the use of sex products (such as toys), or is blatantly erotic or sensual (e.g., strip teases)
- Violent and graphic content: Content that can cause discomfort, shock or disgust to viewers may be ineligible for the FYF, including scary effects, jump scares, makeup that realistically replicates gory wounds, or depictions of bodily functions. Some content may include an 'opt-in' screen or warning
- QR code content: Content that includes QR codes is usually ineligible for the FYF because it can lead users to harmful websites or apps, though TikTok makes exceptions in certain circumstances where that risk is low (e.g., e-commerce).

Prevention of CSEA content or content sexualising children, in 'LIVE' or 'For You' being recommended to adults

eSafety asked what steps TikTok are taking to prevent the recommendation of CSEA content and content involving the sexualisation of children in 'LIVE' or the 'For You' page to adults with interests in similar accounts or content. TikTok stated that its safety features are designed to ensure that CSEA content and content involving the sexualisation of children 'never reaches the LIVE or FYF'. TikTok stated that its zero tolerance policy on CSEA is implemented through its automatic machine detection, user reports and human

moderation capabilities. If CSEA content and content involving the sexualisation of children is shared on 'LIVE' or the FYF, the content is removed, and the violating user is banned. TikTok reported that it takes additional steps that eSafety has decided not to publish to avoid the information being misused.

Additionally, TikTok stated that its recommender system 'works to intersperse recommendations that may fall outside people's expressed preferences... [TikTok's] systems won't recommend two videos in a row made by the same creator or with the same sound.' TikTok stated that it is 'testing ways to avoid recommending a series of similar content to protect against viewing too much of a content category that may be fine as a single video but potentially problematic if viewed in clusters.'

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

TikTok was asked whether it has any internal metrics to assess the efficacy of interventions to detect CSEA. TikTok reported that it does have internal metrics, including:

- Proactive removal rates (i.e. removing a video before it is reported by a user or trusted flagger)
- Number of video removals within 24 hours
- Zero view rate (the rate at which content is removed from the platform before it receives any views)
- Amount of content, and number of reports made to NCMEC

17. Question about appeals against CSEA-related moderation

In response to a question about how many appeals have been made by users for accounts banned or content removed for CSEA material and activity, TikTok responded that 27,575 appeals were received globally during the notice Report Period from users for accounts banned or content removed for CSEA. Of this, 1,566 (5.7%) appeals were successful; and 26,009 (94.3%) appeals were rejected.

TikTok stated that its data retention practices require that it only retains CSEA-related appeals data for 90 days.

TikTok added that:

The true data is likely to be less than the figures stated above, as the above figures include some non-CSEA bans that could not be filtered out. We have filtered out known non-CSEA bans, however, due to the scale of the data and time period to respond to this request, there will be some non-CSEA related cases included in these figures. By way of example, not all of the successful appeals (5.7%) were CSEA related.

18. Questions about Safety by Design training for staff

TikTok reported that it provides training to its staff on Safety by Design principles. TikTok stated:

TikTok has a team of experts that are focused on Minor Safety by design. These experts' role is to ensure safety by design practices are embedded into the service and our policies. Their main responsibilities are to design product principles and policies, develop risk assessments and collaborate with product experts to co-design solutions. There are additional efforts to educate staff on diverse key areas such as minors and privacy, safety by design, and age appropriate design. results.'

TikTok added that human moderators review content which garners larger video views, content that is flagged by machine models, or content that is reported by users. TikTok added that depending on the nature of the content, human moderators 'may choose to apply several actions on the content, which may include: approving it, marking it as a violation and taking it down and/or banning the video/account, removing the video from users' FYF, or isolating the video from being shown in the general FYF and search results.'

8.4 Twitch Summary

Overview

Twitch Interactive, Inc. was asked about its Twitch service.

1. Questions about known and new CSEA material

Twitch was not asked about the detection of known images, as eSafety understands that there is no ability to send still images on Twitch.

Known CSEA videos

Table A

Service	Hash matching to detect known video
Twitch	No

Twitch stated that it does not use hash matching for videos on its service. Twitch answered that ‘live streamed Twitch content is almost always “new”... [so] the resulting stream hashing would be very unlikely to match the hash of a source video.’

In response to a question about any alternative steps Twitch is taking to detect known CSEA video, Twitch stated that it is investing and focusing attention on detecting livestreamed CSEA materials. Twitch noted:

These measures would detect both new and known materials with equal efficacy because they are targeted at the material itself rather than a hash which would not be likely to match the material as it appears on Twitch.

Twitch also stated that only pre-vetted users can upload video content to the service and that all uploaded videos are reviewed to ensure they meet safety standards and Community Guidelines.

New or unknown CSEA detected proactively

In response to questions about whether Twitch uses any tools to detect new CSEA material, Twitch responded that it uses the following tools:

Table B

Service	Tools used to detect new CSEA?	Names of tools used	Description
Twitch	Yes	Stream Scanner	Twitch-built harm detection service used to scan livestream images against computer vision models, primarily AWS Rekognition ¹¹¹ for nudity and sexual content detection.
		Crisp	Crisp proprietary and open-source AI technology to detect, report, and/or remove harmful content including CSEA which may appear on Twitch. ¹¹²
		Groomer Detection Tools	Twitch-developed automated heuristics to identify users who may intend to engage in grooming, and tools to report such users along with potential underage accounts.

Twitch also stated that it uses reporting tools and prioritized escalation to detect new ‘unknown’ CSEA, highlighting that reports of harmful content (including user reports, moderation staff and external engagements) are made through tools which quickly identify and isolate potentially harmful content for further review. Reports of severe violations (including CSEA) are escalated for ‘highly prioritised’ human review. Twitch reported that while it is assessing severe violations a channel may be temporarily hidden..

Twitch stated that human moderators review all reports flagged as CSEA material by these tools.

In response to a question about the steps taken when it is confirmed that an account has shared new CSEA material, Twitch responded that the first priority is to remove access to the material itself. Twitch reported that in many cases, if the material included nudity or was otherwise clearly CSEA, the stream will have already been ended (or the video-on-demand content will have already been removed) by ‘first-responder moderators’. In addition, a ‘strike’ is applied to the account to suspend it. These strikes also lead to the application of measures to combat ban evasion.

¹¹¹ An image and video recognition tool.

¹¹² An organisation that helps identify bad actors and potential victims, and general CSEA risk. Following this initial response, Twitch subsequently provided eSafety with further information regarding tools used to detect new CSEA, and stated, ‘Twitch is constantly investing in, measuring, and optimizing its overall safety and content moderations programs. As a result, the precise constituent tools of those programs and our reliance on those tools is frequently in flux. For example, while Twitch relied in several areas on the third party Crisp at the time of its submission, Twitch has since significantly scaled down its reliance.’

Twitch stated that confirmed CSEA cases are then handled by Twitch’s dedicated Law Enforcement Response (LER) team, which first collects and segregates all relevant data in the CSEA activity if there is a need to involve law enforcement. Where relevant, LER files a National Centre for Missing and Exploited Children (NCMEC) CyberTip report, including both those reports mandated by law and voluntary reports in appropriate cases.

After the immediate actions arising from CSEA detection have taken place, Twitch stated that:

LER engages in a comprehensive investigation of the activity, the actors, and their potential contacts. This may include on- and off-Twitch activity, with the goal of using a single identified event to identify all possible linked actors or actions. Throughout this process, LER continues to engage with NCMEC as necessary, and participates in efforts such as Signal Sharing through the Tech Coalition, as well as other aligned organizations throughout the world, such as INHOPE and ICMEC.

URL blocking

Twitch was asked whether it blocks URLs linking to known CSEA material and provided the following information.

Table C

Part of service	URLs linking to known CSEA material blocked?	URL sources
All parts of the service, including direct messages (“Whispers”)	Yes	Internal investigations and language analysis Crisp ¹¹³ Cross-industry sharing of URLs ¹¹⁴

Other steps are taken by Twitch that eSafety has decided not to publish to avoid the information being misused.

¹¹³ Refer to previous footnote regarding Twitch's changed reliance on Crisp.
¹¹⁴ Through ‘Threatexchange’, Tech Coalition, Thorn, and others.

Proportion of CSEA material detected proactively

In answer to what proportion of CSEA material is detected proactively, as compared to CSEA material reported by users/trusted flaggers, Twitch responded that it does not track this metric as it often receives multiple reports (both from proactive detection and those made by users) for one instance of CSEA material.

Following this initial response, Twitch subsequently provided eSafety with further information regarding proactive detection, and stated:

Since receiving this request for information from eSafety, Twitch has worked to further investigate this number and develop more accurate metrics responsive to eSafety's question. We can now answer that in Q1 2023, ~20% of CSEA material was identified via proactive detection systems and proactive investigation by Twitch's safety team. Twitch clarified that this figure does not include the prevention of material occurring on the platform in the first place.

Screen-capturing detection

eSafety noted that screen-capturing is a common means of disseminating illegal content, including CSEA content, for sharing on other platforms. Twitch was asked whether it had been working on any measures to detect screen-capturing, or attribute screen-capturing during the Report Period.

Twitch responded 'no'.

Twitch added that:

because screen-capturing occurs on the user's host device with third-party software, Twitch is unaware of an effective means to detect screen capturing, nor the potential to develop such a means through increased investment.

Twitch stated that where the capture was made using Twitch's Clips/Highlights features (i.e. using on-site capturing tools), Twitch enforces its prohibition of CSEA content against these video artifacts. Twitch noted that it combats screen-capturing by collaborating with industry organisations such as Thorn, the Tech Coalition and INHOPE to share information about threats, perpetrators and effective techniques to counter CSEA dissemination.

Twitch added that:

These efforts, in combination with Twitch's work to investigate off-site conduct and take action based on off-site CSEA activity... and the measures Twitch employs to prevent child exploitation from occurring are the most effective ways to prevent CSEA dissemination in the context of screen-capturing, where the action itself is taken on an end-user device.

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

Twitch was asked whether it reviews an account for other CSEA material after it confirms that the account has shared CSEA material. Twitch responded that once it confirms that an account has shared CSEA material, its LER team commences an investigation, as outlined. Twitch stated that it prioritises its investigation resources according to egregiousness, risk, and use patterns.

In some cases, Twitch reported that it reviews a subset of other material shared by the account, though its LER team's investigation may involve a review of all content. Twitch added that technical tools are also used depending on the circumstances of the investigation, and that internal tools such as an internal enforcement tool and Toolkit are used as well as external tools, such as Crisp¹¹⁵.

3. Questions about measures in place to detect the livestreaming of child sexual exploitation and abuse on Twitch

Detection of CSEA material that is livestreamed

Twitch was asked if it had any measures in place to detect the livestreaming of CSEA on its service. Twitch responded that it does have measures in place and uses the following tools, as set outlined the table below:

¹¹⁵ Refer to previous footnote regarding Twitch's changed reliance on Crisp.

Table D

Tool	Description
Language analysis tools (text)	Detection of harmful text including text indicative of CSEA, which eSafety has decided not to publish, and URLs, to prevent the use of language or links indicative of CSEA from being used in chat, Whispers, bios, usernames, and any other text entry fields on Twitch services. Blocking certain search terms associated with CSEA. Twitch also uses Spirit AI.
Age assessment tools	Tools to scan account information entered by users that might suggest an account is operated by a child. These accounts are then flagged for suspension.
Video classifiers	Stream Scanner uses AWS Rekognition to scan images from video streams for evidence of nudity. Detected nudity is flagged for review by human moderators.
Behavioural indicators	Account verification to a phone number before users can create streams in categories that are often streamed or viewed by youth. Twitch reported that this adds a layer of protection for accounts operated by minors who have parental monitoring tied to their devices, making it harder for a minor to create a livestream. Twitch also stated that it uses behavioural tools to detect individuals suspected of engaging in grooming activities and users who may present a risk of CSEA activity based on a risk model. Twitch stated that in the fourth quarter of 2022, the model generated 318 average daily reports for suspected online grooming with 20% of those reports leading to user suspensions, and 37% of these user suspensions were marked under the “severe” suspension reason for review by the LER team.
Traffic indicators	A Twitch model which reports streamers (as suspected underage accounts) and chatters (as suspected perpetrators) for moderation review based on traffic from certain referrer sites. Reported accounts are added to a list of suspicious referrers.
Law enforcement response team and off-service conduct investigations	<p>LER team investigations can identify practices and patterns which inform safety tooling and automated intervention projects.</p> <p>‘Because of Twitch’s live-oriented content, we look at users’ conduct comprehensively, rather than waiting until a potential harmful act takes place on Twitch itself. Twitch investigates reports with evidence of off-service CSEA (and other behaviours) and, if able to confirm, issues enforcements against the relevant users. We apply our Off-Service Conduct Policy standards even if the off-service target of these behaviours is not a Twitch user and has not yet engaged with Twitch, as an extra layer of protection.’</p> <p>Twitch stated that users engaging in severe offenses off-services are also prohibited from subsequently registering Twitch accounts.</p>

Tool	Description
Third-party monitoring	Engaging third parties, such as Crisp, ¹¹⁶ for detection methods and protocols, and reporting indications of activity to Twitch. Twitch also stated that it engages teams of human moderators to review content and act directly on egregious harms. 'One important benefit of Twitch's investment in this area is our ability to review all video content that is uploaded (rather than live streamed) by accounts which have permission to do so.'
Cross-industry group participation	Twitch stated the LER team works with Crisp, ¹¹⁷ the Tech Coalition, Thorn and others, to share learnings and information.

Twitch stated that human moderators view all reports of livestreamed CSEA.

4. Questions about language analysis technology used to detect CSEA activity

Language analysis technology to detect likely grooming

Twitch was asked if it uses language analysis technology to detect grooming. Twitch stated that it does use tools on both public chat and Whispers.

Twitch reported that the language analysis tools used on public chat include AutoMod, tools used by Crisp,¹¹⁸ and other tools which eSafety has decided not to publish. Twitch also reported using language analysis tools on Whispers which eSafety has decided not to publish.

Twitch added that human moderators review all reports flagged by its language analysis indicators.

In response to a question about what steps Twitch takes when grooming indicators are detected on an account, Twitch reported that it takes the same steps as those taken when it confirms an account has shared new CSEA material (i.e. any CSEA content is removed, and the LER team commences an investigation).

¹¹⁶ Refer to previous footnote regarding Twitch's changed reliance on Crisp.

¹¹⁷ Refer to previous footnote regarding Twitch's changed reliance on Crisp.

¹¹⁸ Refer to previous footnote regarding Twitch's changed reliance on Crisp.

Language analysis technology to detect terms, abbreviations, codes and hashtags indicating likely CSEA activity, in particular but not limited to sexual extortion

In answer to a question about whether Twitch uses language analysis technology to detect terms, abbreviations, codes and hashtags indicating likely CSEA activity, in particular but not limited to, sexual extortion, Twitch reported that they were used on both public chat and Whispers, as set out in section 4 above.

Twitch reported that its specialised internal teams and Amazon teams conduct research to identify terms relating to CSEA and that it also leverages research from organisations such as Thorn to identify and update terms known to be used for CSEA. Twitch also stated that human moderators review all reports flagged by these indicators. In response to a question about what steps Twitch takes when these indicators are detected on an account, Twitch reported that it takes the same steps as those taken when it confirms an account has shared new CSEA material.

5. Question about information collected from users on sign up

Information collected

In response to a question about what user identifying information is collected on sign up to Twitch, Twitch responded that the following information is collected:

- Phone number or email address
- Device identifiers
- IP address
- Chosen username
- Date of birth
- Password

Authentication of accounts

In response to a question about whether there is an authentication process for any of the information captured on sign-up to ensure the account is controlled by the user signing up, Twitch confirmed that a verification prompt sent to either the provided phone number or email address authenticates the information captured on sign-up to ensure the account is controlled by the user signing up. Twitch reported additional steps that may be taken which eSafety has decided not to publish.

6. Question about steps to prevent recidivism

In response to a question about what measures Twitch has in place to prevent recidivism of users for CSEA-related reasons on its service, Twitch responded that it uses several¹¹⁹ indicators, which eSafety has chosen not to publish, to prevent the information from being misused.

Twitch was asked, if any information collected on sign up (outlined in section 5 above) was not used to detect recidivism, to provide reasons why not. Twitch explained that some categories of information are not used, which eSafety has decided not to publish to avoid the information being misused.

7. Questions about reporting in relation to CSEA material and activity

'In-service' reporting

Twitch was asked if users can report instances of CSEA activity or material to Twitch within the service (as opposed to navigating to a separate webform or email address).

Twitch stated that users who are signed in to an account are able to make 'in-service' reports but that this is not available to users who are not signed in to an account. Twitch stated that its reporting infrastructure was designed to ensure reports can be attributed to a specific user and to 'mitigate the risk of automated mass reporting by bots or other malicious actors', which streamlines the responses to reports.

Twitch stated that it is working on a way to allow users who are not signed in to submit reports, and that it will monitor its efficacy in responding to user reports. Twitch did not specify when this trial will conclude.

Twitch explained that it has alternative methods by which users of the service, irrespective of whether they have an account, can submit reports of CSEA material including emailing Twitch support or contacting the Twitch Support Twitter account.

Twitch reported that users who are not signed in to an account cannot report instances of CSEA material or activity using a specific reporting category for CSEA. Twitch stated that it

¹¹⁹ eSafety uses the following terms to give an impression of the extent of the indicators used by services in the table below, rather than publishing the specific indicators which could be misused:

- Minimal: A small number
- Several: A moderate number
- Multiple: A significant number.

utilises tools to prioritise user reports once received. Twitch explained that these reports are then assessed as to whether they are more likely to contain CSEA so that they can be sent immediately to the appropriate team and addressed as a high priority. However, Twitch re-stated that there are benefits to requiring users to be logged in before they can submit a report and that its reporting system was designed with that policy in mind.

Median time for a user report of CSEA to be actioned

Twitch was asked to provide the median time taken to respond¹²⁰ to a user report about CSEA:

Table E

Services	Median time for user reported CSEA material to be actioned
Twitch	<p>8.22 minutes¹²¹</p> <p>Twitch stated that 8.22 minutes represents the total amount of time between when a report is first made, and an enforcement action being taken against the account holder.</p> <p>Twitch stated that by the time a report of CSEA is reviewed, if the content involved nudity or clearly CSEA video/images (as opposed to text-based CSEA such as grooming), it likely has already been removed by moderation staff. Twitch also stated that the time reflects enforcement against the account-holder, not actions taken on the content itself.</p>

Following a subsequent question from eSafety, seeking clarification of Twitch's response to the Notice with regards to which point in the reporting process the content is removed. Twitch answered that:

as they are triaging a report, initial and intermediary reviewers are able to immediately stop an in-progress stream or remove a stored VOD if it appears to be in violation of Twitch community guidelines, then escalate for further review as to the appropriate enforcement action...content which is clearly harmful can be removed from the service as quickly as possible, while allowing [Twitch] moderation staff to assess and apply the appropriate enforcement against the streamer.

¹²⁰ Defined in the Notice as “Content removed and reported, user banned, or other content moderation decision made.”
¹²¹ For users signed into the service

8. Questions about steps to identify and remove underage users

Tools to detect underage users

In response to a question about what tools Twitch uses to detect underage users, Twitch stated that it uses the following tools:

- **Age entry upon sign-up** and that ‘Users are unable to immediately enter a different date of birth and are simply prompted that they are ineligible to make an account’ if they attempt to sign up after entering an ineligible date of birth.
- **Text analysis** to scan text entered by users in various portions of the site to look for statements which indicate users are underage.
- **Traffic analysis** based on suspicious referrer URLs to identify and report channels which may belong to underage users based on the traffic they receive.
- **Parental report submission** so that parents whose underage children may be using Twitch services can make a report, which is handled through a dedicated workflow and results in the suspension of the offending account when confirmed.
- **Report evaluation and analysis** where moderation teams review not only the evidence of the reported violation, but also whether the user in question may be underage. In this case, the report is escalated for further investigation.

In answer to a question about how Twitch determines if the reported account is indeed underage, Twitch responded that ‘if a channel is reported by users or internal tools as belonging to an underage user, the Twitch moderation team makes an assessment of the user’s approximate age’ using a range of indicators and then takes action on an account if the user reasonably appears to be younger than 13. These account holders may appeal the process and provide evidence of age.

Number of underage users identified

In response to a question about the number of underage users Twitch has identified and removed during the Report Period, Twitch advised that it identified and removed the following number of underage users:

Table F

Region	Number of underage users removed	Proportion detected proactively
Globally	120,070 users	14.8%
Australia	2,590 users	14.4%

Twitch also stated that a significant contribution to its proactive detection proportion is due to ‘youth who are unable to create a new account due to proactive detection methods applied to users attempting to create a new account, who were previously removed for being underage.’ This contribution was not included in the numbers provided by Twitch.

Twitch reported that it has implemented various methods to prevent users who have been removed from the service for violating the Community Guidelines or Terms of Service from opening new Twitch accounts.

In addition, Twitch noted that:

As an additional measure intended to prevent underage users from using Twitch’s service, the Twitch mobile app is rated “17+” on the Apple App store and “Teen” on the Android store. This is designed to prevent children from installing the Twitch app based on controls made available within those marketplaces.

ID requests

eSafety noted that Twitch can request ID if a user is suspended for being underage and subsequently appeals that suspension. During the Report Period, Twitch stated that it had requested ID from a suspected underage user to confirm their age on 202 occasions globally and that 6 of these were from Australia.

9. Questions about human moderation

Channel creators’ reporting obligations for CSEA material

eSafety stated in the Notice that it understood that that channel creators, streamers and moderators have the ability to remove and ban users within a channel or stream. In answer to a question about whether Trust and Safety staff are advised when channel creator/streamers or volunteer moderators issue a temporary and/or permanent ban on a user for CSEA related activity in a channel, Twitch responded that Trust and Safety staff are not advised of the CSEA breach.

Twitch stated that it takes a ‘layered’ approach to safety ‘which balances consistent sitewide safety standards with more personalization at the channel level.’

Following this initial response, Twitch subsequently provided eSafety with further information regarding human moderation, and stated:

Twitch Trust and Safety staff are able to review user content, and do so whenever a report of such content is received or when such content is detected or suspected through our proactive tools... channel-level moderation and “bans” are not associated with any reason—they are intended to be flexible so that they can reflect a channel’s own rules and norms, which are on top of the baseline Twitch-wide Community Guidelines. Therefore, there is no ban reason created for any channel-level moderation activity done by streamer and volunteer moderators. For this reason, the path for any breach of Twitch-wide Community Guidelines, rather than channel-specific rules, is to file a report with Twitch Trust & Safety rather than to use channel moderation tools.

In answer to a question about how Twitch ensures that a reported user who is banned on one part of its service is not able to engage in CSEA related activity on other parts of the service, Twitch stated that ‘in addition to [its] Trust & Safety teams, tools, and resources, every user of Twitch—streamers, moderators, and viewers—are encouraged to use [its] reporting tools to report violations. If these reports result in [its] team identifying a violation, the user shall be prohibited from accessing the service as a whole, not just a particular channel.’

In addition, Twitch noted that CSEA violations are reported using site-wide tools as opposed to channel moderation. Twitch stated that it ‘has not seen any evidence that CSEA goes unreported due to the nature of this layered approach...in 2022 Twitch launched a tool (“Shared Ban Info”) that allows streamers to share information about who they’ve banned in their channels with one another.’

Support and guidance for creators/streamers and volunteer moderators in relation to CSEA

In answer to a question about what support or guidance Twitch provides to channel creators/streamers and volunteer moderators in relation to ensuring CSEA is proactively minimised and accounts removed, Twitch responded that it provides a Safety Center on its website with educational material including information on CSEA material and conduct. Twitch stated that the Safety Center encourages users to create reports and resources to teach effective reporting and moderation.

Twitch provided information about the following tools which it explained have particular features that relate specifically to CSEA material:

Table G

Tool name	Purpose
Automod	AutoMod can be configured to block or hold from chat a wide variety of harmful terms which can be associated with CSEA.
Blocked Terms	Blocked Terms allows streamers to add additional words which cannot appear in their chat.
Shield Mode	Shield Mode is a one-click solution to change a variety of channel moderation settings quickly, particularly when a user perceives a sudden influx of harassment or other inappropriate activity (including CSEA). While the specifics can be configured by the user, Shield Mode can, for example, prevent all users from chatting.

In answer to other questions on support provided to creators/streamers and voluntary moderators to help them respond to CSEA, Twitch stated that it provides streamers with tools to decrease the potential for CSEA to appear on their channel ‘for example by setting their channel chat to allow only those users who are subscribers to their channel, or only those users who have a phone number-verified account.’

A standard policy, or similar, outlining the responsibilities and expectations of the channel creator/streamer and volunteer moderator roles

In answer to a question about whether Twitch has a standard policy, or similar, which outlines the responsibilities and expectations of the creator/streamer and volunteer moderator roles, Twitch responded that advice and policies are available on the Safety Centre. Twitch added that guidelines on the Safety Centre clarify the expectations of streamers and how they should use moderation to ensure safety, stating that streamers ‘are required to use moderators and tools to mitigate inappropriate conduct.’ Twitch responded that it also has an internal policy relating to the responsibilities and expectations of content moderation personnel.

Twitch explained that if a channel creator or moderator does not meet the standards, they could face a warning or suspension of various lengths depending on the severity of the conduct.

User's ability to make reports about channel creators/streamers and volunteer moderators

In response to a question about whether users are able to make reports about channel creators/streamers and volunteer moderators in instances where they are failing to effectively moderate, Twitch answered yes, users can.

However, Twitch outlined that the only way a user can report a channel creator/streamer or moderator is to choose the category that applies most appropriately to the content that is violating the Community Guidelines.

Twitch stated that it does not track how many reports are made about channel creators/streamers and moderators. However, Twitch stated 'that a suspension for "Unmoderated Hateful Conduct or Harassment" was applied to 7,312 creators during the Report Period.'

In response to a follow up question from eSafety, Twitch stated that the 'Unmoderated Hateful Conduct or Harassment' is a category assigned internally by Twitch rather than a category that users can choose when reporting channel creators/streamers and moderators for failing to effectively moderate. Twitch added that:

The notion behind this is that users are encouraged to report conduct rather than individuals, which allows our moderation teams to collect more accurate information for their investigation and ascribe responsibility for the violating conduct as necessary.

Twitch also stated that its moderation team (and, in the case of CSEA, its LER team) will review the report and 'determine whether the proliferation of the content was due to an individual chatter's acts, and/or whether the channel owner is also culpable or deserving of enforcement.'

10. Question about what languages human moderators operate across

eSafety asked Twitch to list all the languages that Twitch's human moderators (employees and contractors) operate across. Twitch listed the following languages:

Table H

Amongst full-time employees of Twitch, the following languages are covered:			
Arabic	Chinese (Cantonese)	Chinese (Mandarin)	Chinese (Taiwanese)
Danish	Dutch	French	German
Indonesian	Italian	Japanese	Korean
Malay	Norwegian	Polish	Portuguese
Russian	Spanish	Swedish	Thai
Turkish			
Twitch's third-party vendors also provide coverage in the following languages:			
Arabic	Chinese (Cantonese)	Chinese (Mandarin)	Czech
Dutch	English	French	German
Italian	Japanese	Korean	Polish
Portuguese	Russian	Spanish	Swedish
Thai	Turkish	Vietnamese	
Twitch explained that the following languages are either languages that will be implemented (with the addition of an additional vendor to cover event moderation) or 'can be implemented' upon identification of a need:			
Bosnian	Bulgarian	Burmese	Catalan
Croatian	Danish	Estonian	Farsi
Finnish	Frisian	Greek	Gujarati
Hebrew	Hindi	Hungarian	Indonesian
Khmer	Latvian	Macedonian	Malay
Montenegrin	Norwegian	Romanian	Serbian
Slovak	Tagalog	Tamazight	Tamil
Ukrainian	Urdu	Welsh	

Following a subsequent question from eSafety, seeking clarification of Twitch's response to the Notice regarding what it meant by languages which can be implemented 'upon identification of a need', Twitch answered that it considers the combination of viewership data for a language, and volumes of reports in a language, over the preceding three months to either increase or decrease language coverage. Twitch also noted that 'there is a process to account for cases in which [Twitch] anticipate[s] a need to increase support for a

particular language, such as an event that will cause a marked increase in viewership for that language. Note that in instances where a particular language is not yet ‘supported’, user reports and queries are still addressed as necessary through ad hoc translation.’

11. Question about information provided in NCMEC reports

In response to a question about the information Twitch includes in its reports to NCMEC, Twitch provided a list of information that it includes in a report to NCMEC. eSafety has correlated this with the types of information NCMEC defines as ‘actionable’ for law enforcement.

12. Questions about Twitch recommender systems

Design objectives

Twitch stated, in response to a question about the top design objectives of Twitch’s recommender system, that part of its objectives are to present a user with channels that match what users similar to them have watched in the past and the user’s own viewing patterns. eSafety notes that these are not design objectives, rather a means of delivering on the design objectives of recommender systems. Following a subsequent question from eSafety, seeking clarification of Twitch’s response to the Notice regarding what Twitch designed its recommender system to achieve, Twitch answered:

the only design objective for the Twitch recommender system is for connecting viewers to streamers and communities that they can engage and be a part of.

Signals used to recommend content and their importance in meeting the design objectives

Twitch was asked to list and describe the primary signals that it uses in determining what content should be recommended to a user and provide details of their relative importance in meeting the recommender design objectives.

Twitch provided a list of signals, including:

- ‘behavioral data such as watch history and history of users with similar behaviors’.

eSafety has decided not to publish the full response provided by Twitch, which Twitch submitted was commercial in confidence information.

Questions about risk impact assessments and auditing

Twitch was asked questions about the measures it employs to test and update its system to improve the safety of its service and avoid amplifying harmful communities. Twitch stated that ‘when content creators are identified and removed as part of Twitch’s Harmful Misinformation Actors community guidelines, those creators are also removed from recommendations. Twitch believes this removes the scope for amplification of harmful content. [It relies] on the bad actor list and ensure that [its] recommender systems check against this list before recommending any content. Suspended users are not recommended in any case.’ Twitch also noted that for marketing purposes, creator accounts who have received recent strikes may be removed from the list of recommended channels.

In response to a question about whether tests are performed on their recommender systems to ensure the overall safety of Twitch’s platform and the frequency of those tests, Twitch responded that it performs ‘periodic performance tests and engineering updates as and when necessary.’

Age-appropriate restrictions to recommendations

Twitch was asked to describe the criteria it uses to evaluate content that is harmful to children but does not violate community guidelines and any approaches it takes to not recommend or recommend less often.

Twitch responded that its primary focus is on age restrictions rather than outright removal. It uses content classification labels for content which does not violate its Community Guidelines but which may not be suitable for users under 18. A system is currently under development and slated for launch in the second half of this year, where, once launched, Australian users under 18 will be unable to access content classified as including certain mature subject matter.

Twitch later confirmed that this system has been released and is operating in Australia.

Following a subsequent question from eSafety, seeking clarification of Twitch's response to the Notice about steps taken during the Report Period, Twitch also stated that:

it is important to note that Twitch's Community Guidelines, in language and in application, are more stringent and set a higher bar than many other user-generated content services. Nonetheless, Twitch recognizes that some content may not be appropriate for all users. With this in mind, during the reporting period, Twitch made use of the "Mature Flag," which is a setting streamers may apply if their content is not appropriate for some audiences. Viewers attempting to open the stream would see a warning about the potentially mature content, which must be acknowledged before being able to view the stream.

Mitigations utilised to minimise the risks associated with recommender systems promoting children's accounts to adults who are seeking to groom, or sexually exploit them

Twitch was asked what mitigations it uses to minimise the risks associated with recommender systems promoting children's accounts to adults who are seeking to groom, or sexually exploit them. Twitch responded that it uses several mitigations to minimise the risks associated with recommender systems including search-term blocking to prevent potential groomers from using terms and phrases to identify vulnerable streamer channels. Twitch added that it limits the sorting channel options so as to limit how easy it is for users to search for vulnerable streamer channels.

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

Twitch was asked if it has internal metrics to assess the efficacy of its intervention to detect CSEA. Twitch responded that it does have internal metrics, including:

- Regularly reviewing cases that are escalated by Twitch's third-party vendor (who carries out the first round of review on all user reports that are received by Twitch, and on potentially violative content that is proactively detected by automated systems). This review is to assess the effectiveness and efficacy of the vendor's assessment by determining how many of those cases are found to be false-positive cases. Twitch also stated that it carries out random sampling of cases that are not escalated by the vendor, as part of a quality assurance process, to determine how many cases are found to be false negative (i.e. how many of the cases should have been escalated but were not).

- Actively tracking cases that are escalated and/or where enforcement action is taken and matching these against product details (such as where on Twitch the incident took place). Twitch states this allows it to determine high-risk product features, and feedback is then provided to the product/engineering teams as appropriate.
- Receiving monthly reports from NCMEC regarding the number of cases Twitch submitted which led to investigations being opened by law enforcement agencies globally.

14. Question about appeals against CSEA-related moderation

Twitch was asked to provide the number of appeals made by users that had been banned or had content removed for CSEA-related violations. Twitch responded that 7,570 appeals had been made by users in the Report Period for content/accounts removed for CSEA material or activity globally, and that 1,255 were successful. At the date of Twitch's response, it noted that 1,365 of the appeals had been rejected and the remaining 4,950 remained 'open'.

Twitch stated that appeals were only granted if the person appealing did not carry out any violative conduct (i.e. based on reported conduct or conduct discovered through investigation), or if the context provided by the person appealing is found to be relevant.

15. Questions about Safety by Design training for staff

In response to whether Twitch provides training to its staff on Safety by Design principles, Twitch responded that it does and that a Safety by Design assessment is in place for new products that are being developed. Additionally, Twitch added:

On March 10, 2022, the Internet Commission, a group of independent experts whose mission is to advance digital responsibility through independent evaluation, released its Accountability Report 2.0. This report evaluated the online trust and safety practices of multiple organizations, including Twitch. The report describes, among other things, how "Twitch has integrated safety by design," including by "training project managers in safety, requiring all specifications to be reviewed to identify possible safety risks, setting positive expectations for user safety, and aligning safety priorities across functional teams." The report recognizes that "[e]mbedding multi-stakeholder review into the product development process demonstrates Twitch's commitment to ensuring a sense of shared responsibility for the safety and wellbeing of users across the organization."

8.5 Discord Summary

Overview

Discord Inc. was asked about its Discord service.

Note: eSafety used the term child sexual exploitation and abuse (CSEA) in the Notice given to Discord. However, Discord’s responses sometimes used the term CSAM (referring to child sexual abuse material), rather than CSEA.

1. Questions about known and new CSEA material

Known CSEA images

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

Table A

Service	Hash matching to detect known images?	Names of tools
Direct messages	Yes	PhotoDNA
Servers (public and private)	Yes	Contrastive Language- Image Pretraining (CLIP) ¹²²
Profile pictures (avatars)	No	

Discord stated that in order to ‘maximize the utility and minimize the rate of false positives’ it sources vetted hashes from the US based National Centre for Missing and Exploited Children’s (NCMEC) Non-Profit Organisation CSAM Hash-Sharing Initiative.

When asked how often it updates its hash lists from this database, Discord stated:

Discord updates its hashes against the NCMEC database approximately once per year.

Discord also reported that it maintains its own internal hash database. Discord explained that it uses CLIP to identify new ‘unknown’ CSEA material which it then hashes and adds to its own internal hash database, and reports to NCMEC. Discord stated that all images

¹²² AI model developed by OpenAI. eSafety notes this is not a hash matching tool.

flagged by CLIP are reviewed and verified by human moderators and that its internal list ‘updates whenever new unknown CSEA material is detected by CLIP.’

Discord explained that it does not scan avatars (profile pictures) for known CSEA images as it does not ‘consider there to be a material risk of CSEA images being uploaded as user avatars.’ Discord explained this is ‘due, at least in part, to the public-facing and conspicuous nature of avatar images’ and because ‘avatars tend to be manipulated...versions of full images which decreases the efficacy of industry hashes.’ Discord reported that it ‘aims to focus its technical efforts and human resources proportionally towards active risks that exist on [its] services.’ Discord noted that users can report inappropriate or violative avatars.

Known CSEA videos

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

Table B

Service	Hash matching to detect known video?
Direct messages	No
Servers (public and private)	No

Discord stated that it does not use tools to scan for known CSEA videos on any part of its service. Discord reported that this is due to the ‘prioritization of image-based content on the service and the higher resource burden associated with hash matching on video material’ also, that ‘video is often shared as links or embedded videos from other platforms, subject to their own terms of service and safety monitoring activities.’ Discord noted that it is ‘always exploring ways to efficiently increase our capabilities in this space.’

New or ‘unknown’ CSEA material

In response to questions about whether Discord uses any tools to detect new CSEA material, Discord responded:

Table C

Service	Tools used to detect new CSEA?	Names of tools used
Direct messages	Yes	CLIP, which Discord described as ‘Discord’s implementation of OpenAI’s algorithm.
Servers (public and private)	Yes	

Discord reported that all material detected by CLIP is reviewed by human moderators.

Discord explained that when it detects new or known CSEA material it removes the offending material, bans the offending account and files a report with NCMEC.

URL blocking

Discord stated that it blocks URLs linking to known CSEA material on the following services:

Table D

Parts of service	URLs linking to known CSEA material blocked?
Direct messages	No
Servers (public and private)	No

Discord explained that it does engage in ‘limited URL blocking based on third party providers aiding in the identification of violative communities on Discord, as well as certain known malicious links generally related to malware or phishing.’

Discord reported:

Discord does not deploy URL blocking more generally as the focus of our resources have been directed toward removing bad-actors and violative spaces on our platform that would distribute high harm content, including CSEA material. To this end, Discord enables users to report suspicious or violative links in messages on the platform.

Proportion of CSEA detected proactively

For the Report Period, Discord provided statistics for the proportion of CSEA material it detected proactively (as opposed to being reported by users/trusted flaggers) on the following parts of its service:

Table E

Part of service	Proportion of material detected proactively by Discord (%)
Direct Messages	72.1%
Servers (Public)	87.8%
Servers (Private)	44.5%

Discord stated that in the latter half of 2022 it improved its proactive removal rate for ‘CSAM servers’ to 99%. eSafety notes that the question was not limited to servers dedicated to CSAM, but to all CSEA material on the service.

Alternative steps taken to proactively detect CSEA

In response to several questions asking about the alternative reasonable steps Discord is taking to detect CSEA on its service, Discord stated that it ‘takes CSEA material seriously’ and as a ‘mid-sized company that has grown exponentially over the last several years’ it has had to ‘rigorously prioritize’ its safety-related resourcing.

Discord also outlined a number of actions it takes in addition to its use of hash-matching tools, including using internal tools (such as machine learning techniques) and working with industry partners to detect CSEA distribution networks. Discord reported that these industry partners include ‘peer companies, non-profits and researchers’. Discord added that it also ‘relies on user reporting of violative material as part of our response to CSEA material across the platform.’

Discord also responded that it ‘works with leading groups and partner organizations to assist in the detection of CSEA material and the improvement of our internal policies and processes’. In addition to NCMEC, Discord named the Family Online Safety Institute and the Technology Coalition and noted that it is ‘a frequent sponsor of events dedicated to increasing awareness of, and action on, child safety issues’.

Discord noted that it had partnered with Thorn’s ‘NoFiltr’ to develop online safety resources for young people and provided the following link which directs parents to educational

material to help them understand Discord and how to help their children have positive and safe experiences on Discord's services.

Discord also described a number of initiatives it has taken in recent years to support its response to CSEA:

- 'Greatly investing' in its Trust and Safety teams, including creating a specialised group to respond to CSEA on the service. Discord described the role of the group as being to proactively and reactively respond to this kind of material.
- Building and formalising a policy team including platform policy specialists and a teen safety policy manager who, over the period of a year, updated and expanded Discord's child safety policies.
- Incorporate safety-by-design at the start of the product development cycle through to post-launch monitoring.
- 'Investing deeply' in product approach to combatting CSEA, including formalising a dedicated 'visual safety product team' and a teen safety product team to do more detection and prevention of CSEA.

Discord stated that its focus for 2023 and 2024 was to implement 'additional tooling that will allow us to leverage our approach to improve scalability and efficiency of this work'.

2. Question about steps taken to identify whether an account that is found to have shared CSEA material also reviews other material shared by that account

Discord was asked whether it reviews an account for other CSEA material after it confirms that the account has shared CSEA material. Discord responded that when it confirms that an account has shared CSEA material, 'a Discord agent' will review a 'subset of further content' shared by that account. Discord explained that it reviews 'material which provides context for reported material, such as surrounding messages', and that this process 'involves both human review and the use of internally developed tools designed to provide context to reported text material as well as obscure or obfuscate harmful images for the well-being of Discord's agents.'

Discord reported that its teams 'review all reports related to CSEA material' and that in certain instances a CSEA report may be escalated to its Exploitative Content team who do further investigation into context and circumstance.

3. Questions about measures in place to detect the livestreaming of child sexual exploitation and abuse on Discord

Discord was asked if it had any measures in place to detect the livestreaming of CSEA on its service. Discord responded with the following information:

Table F

Services	Measures in place to detect CSEA in livestreams or video calls/conferences	Names of tools used
Discord	No	

Discord explained that it ‘does not monitor or record livestream content or voice chats’ and that it has ‘prioritized resources into other forms of CSEA detection’. Discord stated that:

running models across this type of content at the scale Discord operates would be prohibitively expensive and would operate at the detriment of other Discord safety programs.

4. Questions about language analysis technology used to detect CSEA activity

Language analysis to detect grooming

Discord stated that it does not use any language analysis technologies to detect likely online grooming on its service.

Discord stated that it aims to deploy a ‘grooming classifier’ in 2023 and that it is currently ‘investigating different approaches’ and ‘actively engaged in trials with external parties to develop language analysis technologies to detect grooming’.

Language analysis technology to detect likely CSEA activity, including sexual extortion and the exchange of gifts for underage sexual acts

Discord was asked whether it uses any language analysis tools to detect likely terms, abbreviations, codes or hashtags indicating likely CSEA activity such as sexual extortion or the exchange of gifts for underage sexual acts. Discord responded ‘no’.

Discord stated:

At this time, language analysis tools of the kind described in this question are not deployed by Discord due to the prioritization of technical resources into other CSEA detection and disruption programs.

Discord added that it is currently researching how sexual extortion occurs on its platform and that it has contracted a private firm that specialises in using digital intelligence to detect sexual extortion and trafficking, to increase Discord's awareness of activities such as the sharing, or threat to share, of non-consensual sexual images and sexual extortion. Discord explained that it has developed policies against sexual extortion with its 'recently updated threats policy' which prohibits the threat or intention to share private sexual content without consent from the owner.

5. Question about information collected from users on sign up

In response to a question about what user identifying information is collected on sign up to Discord, Discord responded that it collects username, email address, phone number (if applicable), and other information which eSafety has chosen not to publish to prevent the information from being misused.

In answer to a question about whether it authenticates any of this information, Discord stated it has an authentication process in place to confirm that the email or phone number used to sign up can be accessed by the person who registered using those credentials.

6. Questions about steps to prevent recidivism

Discord reported that a user banned from a server for a CSEA-related violation of its Terms of Service or Community Guidelines is 'banned from accessing their account and therefore from all servers in which they have joined.'

Discord listed multiple¹²³ indicators to detect users that have previously been banned for CSEA-related breaches, which eSafety has chosen not to publish to prevent the information from being misused.

Discord stated that it ‘has a variety of tools at its disposal’ and that it can ‘ban users based on collected identifiers’ to prevent CSEA recidivism on the service.

Discord was asked, if any information collected on sign up (outlined in section 5 above) was not used to detect recidivism, to provide reasons why not. Discord explained why some information was not used, which eSafety has decided not to publish to avoid the information being misused.

7. Questions about reporting in relation to CSEA material and activity

‘In-service’ reporting on different versions of Discord

Discord was asked if users can report instances of CSEA activity or material to Discord within various parts of its service (as opposed to navigating to a separate webform or email address).

Discord stated that there are in-service options that enable users to report instances of CSEA in direct messages and in server content. Discord outlined the sequence of drop-down menu categories that users must select to directly report CSEA using these in-service options.

Discord stated that there are no options for reporting CSEA on server livestreams and audio, except the ability for a user to report chat messages accompanying a live stream.

Following a subsequent question from eSafety, seeking clarification of Discord's response to the Notice, regarding options to report CSEA on server livestreams and audio, Discord responded that there are two options for users to report CSEA on these parts of the service. The first is when a user is on a mobile (either iOS or Android) they are able to report the server itself using the following flow “Report server -> explicit, graphic, or unwanted sexual content -> server promotes sexual content or behavior involving minors -

¹²³ eSafety uses the following terms to give an impression of the extent of the indicators used by services in the table below, rather than publishing the specific indicators which could be misused:

- Minimal: A small number
- Several: A moderate number
- Multiple: A significant number.

> photos or videos depicting real world child sexual abuse”. The second is via a webform on their web portal.

In response to a question about differences in reporting processes for the mobile app, desktop app, and web browser versions of its service Discord initially stated that there were no differences.

eSafety followed up with Discord stating that at the time of asking Discord this question, eSafety’s understanding of the process for reporting violative content and behaviour on the mobile app version of Discord was different from the process for making a report on its web browser and desktop app versions. In this follow-up eSafety sought clarification on Discord's response to the Notice asking Discord to explain why there were different reporting processes on these surfaces.

Discord explained that while in-app reporting was available during the entire Report Period, ‘For desktop and web users, in-app reporting began rolling out during the Reporting Period and was available to some Australian users before the end of the Reporting period.’ Prior to the introduction of in-app reporting on web app and desktop, Discord stated that user reports on these versions ‘would have needed to be submitted via webform’.

Discord added that ‘Following the close of the Reporting Period, in-app reporting rolled out to all users across all platforms and reached 100% of users on or around March 13, 2023.’

Using tools to prioritise reports about CSEA

In response to a question about whether Discord uses tools to prioritise user reports that may contain CSEA material or activity, Discord stated that it does use tools to prioritise such reports adding that reports are ‘routed based on reporting type to Discord moderators specially trained in handling different types of CSEA material.’ Discord stated that reports of CSAM or inappropriate contact with a minor are routed to a ‘high priority child safety queue.’ Following evaluation Discord noted that its moderators can ban users and send a report to NCMEC. Discord also added that ‘in certain instances reports may be escalated to its ‘Exploitative Content team’. Discord explained that this team are able to ‘conduct wider investigations’ and engage law enforcement agencies.

Median time for a user report of CSEA to be actioned

Discord was asked to provide the median time taken by Trust and Safety staff to respond¹²⁴ to a user report about CSEA on various parts of its service:

Table G

Part of service	Median time
Direct messages	13 hours
Servers (public)	8 hours
Servers (private)	6 hours
Server livestreams	Discord stated it was unable to calculate the response time as there is no in-service reporting option.

Following a subsequent question from eSafety, seeking clarification of Discord's response to the Notice and why the question was not applicable to its livestreamed content, Discord responded that because it ‘does not have a mechanism for a user to directly report a server livestream’ it was ‘not reasonably able to calculate a median response time taken by Trust and Safety staff to respond to user reports about CSEA in livestreams’.

Storing content shared by banned accounts

In response to questions about storing content sent or shared by accounts removed for CSEA, Discord highlighted that it retains content shared in direct messages and servers, but not from server livestreams and audio. Discord stated that it does not store video and audio content from voice chats or livestreams and ‘therefore [it] is unable to be archived’. Discord further stated that the ‘resource-intensive nature of storing such material presents a substantial barrier, and Discord instead has chosen to prioritize other means of detection and disruption programs.’ Discord noted that ‘While audio or video content itself is not stored, associated metadata is and can be examined for information where required.’

Discord stated that for direct messages and server content shared by an account banned for CSEA, Discord archives the ‘Datetime, messages, media, participants, event data related to most actions including voice channel joins and leaves’. Discord noted that it stores this information for 270 days.

¹²⁴ Defined in the Notice as “Content removed and reported, user banned, or other content moderation decision made.”

Requiring 'message links' to report abuse via its webform

At the time of issuing the Notice to Discord, eSafety's understanding of the process to report illegal or abusive content using Discord's reporting webform was for users to copy and paste a 'message link' of the specific content they wished to report. Users have previously reported to eSafety that in order to put an end to abuse they have blocked the abuser or deleted the content but, after doing so, Discord has been unable to address the abuse/potential abuser because the complainant can no longer provide the relevant abusive 'message link'.

eSafety asked Discord if it has any means of locating specific communications without a message link. Discord responded that it does have such means, but it is 'incredibly burdensome on Discord staff' and requires an investigatory approach that Discord characterises as 'highly invasive' to the user making the report and other users in the spaces they share.

Discord also highlighted that it considers screenshots to be a 'less reliable indicator of violative behaviour' because they can be altered. For these reasons, Discord stated that 'while technically possible' it is 'unable to reasonably locate messages without a message link identifier'.

Discord was asked whether it has any plans to implement a means to report content using its webform without a message link. Discord stated:

As Discord does not routinely examine image or text data (with the exception of CSAM), nor does Discord index text on its platform, other reporting mechanisms are infeasible, unduly burdensome, and would substantially impose upon the privacy of the reporting party.

Discord stated instead that it allows users to make reports in-app through the in-app reporting feature, as well as through its web portal.

8. Questions about steps taken to identify and remove underage users

Tools to detect underage users

In response to a question about what tools Discord uses to detect underage users, Discord stated that it 'relies on neutral age gates and user reports'. Discord explained that users must provide a birth date when registering a new account. Discord added that if the birth

date indicates that the applicant is under 13 years old the service prevents them from registering and ‘instead presents a notice screen informing the individual that they are “unable to register”’. Discord explained that if the user attempts to register again ‘they are once again shown the “unable to register” Notice without even getting a chance to make up a new birthdate to try to re-register’.

In response to a question about what measures Discord has in place to ensure users whose accounts are closed for being underage do not open another account, Discord stated that ‘subsequent registration attempts on the same device will also be blocked (even if a qualifying birthdate is subsequently entered).’

Number of underage users identified

In response to a question about the number of underage users Discord has identified and removed during the Report Period, Discord advised that it identified and removed the following number of underage users:

- Globally: 224,000 users
- In Australia: 4,813 users

Discord noted for its global figure that ‘underage’ is not consistent across the globe and that in some countries, users can be banned for being 14 or 15 years of age, instead of 13 years as in Australia.

When asked what proportion of underage users are detected proactively, Discord stated that ‘Approximately 98% of underage user detection is responsive to user reports’, which eSafety understands to mean that 2% of underage user detection was detected proactively by Discord.

Discord also highlighted that to determine if a reported account was indeed created by an underage user, its Trust and Safety team determine if it is a ‘credible report, supported by clear and convincing evidence’ (e.g., from a parent), and where the Trust and Safety team deem it appropriate to, they verify the user’s age and disable the account where they believe it necessary.

9. Questions about human moderation

Question about what languages Discord human moderators operate across

eSafety asked Discord to list all the languages that Discord's human moderators (employees and contractors) operate across. Discord listed the following languages:

Table H

German (DE)	Korean (KO)	Turkish (TR)	Croatian (HRV)
French (FR)	Polish (PL)	Thai (TH)	Czech (CS)
Danish (DK)	Russian (RU)	Vietnamese (VI)	English (en-GB)
Dutch (NL)	Spanish - Spain (es-ES)	Traditional Chinese (zh-TW)	Finnish (FI)
Italian (IT)	Swedish (sv-SE)	Simplified Chinese (zh-CN)	Greek (EL)
Japanese (JA)	Brazilian Portuguese (pt-BR)	Bulgarian (BG)	Hindi (HI)
Lithuanian (LI)	Norwegian (NO)	Romanian (RO)	Ukrainian (UK)
Hungarian (HU)			

Trust & Safety Roles

In response to questions about the roles and responsibilities of Discord's Trust and Safety staff and the processes they follow with regards to keeping children safe, Discord provided information on the following roles filled by different Trust and Safety staff:

- Overall creation of strategy and oversight of Trust and Safety internal and outsourced teams.
- Collaboration between Trust and Safety, Engineering, Data Science, and Product.
- Oversight of proactive initiatives in areas of discrimination, extremism, cybercrime, exploitative content, and antispyam.
- Quality assurance of user report responses.
- Oversight of internal wellness and resilience resources for Trust and Safety staff.
- Oversight of Discord's volunteer moderator support programs, including the Discord Moderator Program and Discord Moderation Academy.

In response to questions about the process for directing user reports to Trust and Safety staff, Discord provided the following information:

Table I

Question	Response
Can users report directly to Trust and Safety staff?	Yes. Users can report with in-app reporting through Discord's web portal.
Are Trust and Safety staff automatically notified if a Volunteer Administrator/Moderator removes a user from a server for a CSEA breach?	No.

Volunteer administrators/moderators

Discord responded to questions about the roles and responsibilities of volunteer administrators/moderators with regards to keeping children safe on its service by providing the following information:

Table J

Role	Responsibilities
Volunteer administrator	'Administrators are the people who create Discord servers around specific interests. They establish the rules for participating, can invite people to join, and oversee the health and well-being of their community. They have broad administrative control and can bring in moderators to manage community members. They can also ban or remove members and, if necessary, remove and replace moderators. Administrators also choose moderators to play a vital role in Discord communities.'
Volunteer moderator	'The responsibilities of a moderator might vary, but their overall role is to ensure that their Discord server is a safe, healthy environment for everyone. They can do things like moderate or delete messages, as well as invite, ban, or suspend people who violate the server's rules. The best moderators typically are seasoned and enthusiastic participants in one or more communities.'

Discord provided the following information in response to questions about the processes its volunteer administrators/moderators follow, and the processes Discord has in place to monitor their conduct and uphold moderation standards:

Table K

Question	Response
If Discord's Trust and Safety staff ban a user from a server for a CSEA-related violation, are the relevant volunteer administrator/moderators advised of the breach?	No. Discord stated that it is responsible for taking action against CSEA-related violations of its service policies. If Discord determines that a server's purpose is centred around CSEA material, it will be shut down. Discord stated that in circumstances where the server is not actioned, the volunteer administrator 'and/or all users may be issued a server warning, be reminded of Discord's Community Guidelines, and that further violations of the policy within their community might result in the termination of the community, the Administrator's account, and the users.'
Does Discord collect information about the moderation decision made by volunteer server moderator, such as speed of response to user reports?	No.
Does Discord have a standards policy, or similar, outlining the responsibilities and expectations of the volunteer administrator/moderator role?	No. Discord stated that it provides the 'Discord Moderator Academy' as a 'resource volunteer moderators can use to understand how to better moderate'. Discord also noted that volunteer administrator/moderators are held to the same standards as any user under the Terms of Service and Community Guidelines.
Are users able to make reports about volunteer moderators in instances where they are failing to effectively moderate?	No. Discord stated that it does not have 'dedicated avenues for complaints regarding Administrators or Moderators' and that 'it is not Discord's practice to regulate creators or moderators regarding the effectiveness of their moderation activity.' Discord also noted that server-specific moderation activity covers a wide variety of behaviour and violations of a server's rules may have no relationship to a breach of Discord's community guidelines or Terms of Service.

Further information and context

In response to an opportunity to provide further information and context on how Discord identifies and addresses channel creators and moderators who are failing to effectively

moderate, Discord outlined that community moderation is run by the volunteer Server Administrators and Moderators and that ‘With support from the Discord team, [volunteer administrators/moderators] work to make Discord servers safe, healthy environments for everyone.’ Discord noted, ‘Administrators can also use automated moderation through server settings, or moderation tools and bots developed in-house or by third parties’ and that Discord has ‘published policies describing many of these controls and tools.’

Discord also noted that it provides ‘server owners (sometimes referred to as Server Administrators or “ModMins”) with the option to enable and customize an automated message suppressing tool called “AutoMod”’. Discord explained that Automod ‘allows Server Administrators to customize their approach to detecting, and if so desired, blocking, undesirable and risky messages’ using keyword filters.

Discord also outlined that ‘If Discord learns of activity in communities that violates Discord's policies, Discord will take appropriate action which may include banning all members of a community and shutting down the server.’

10. Question about safeguarding Student Hubs on Discord

Discord lets users create special ‘Student Hubs’ as student-run servers for study groups and clubs. When users sign in to Discord, a pop-notification often appears asking users if ‘you are a current student’ and asking for the user’s school e-mail address. eSafety asked Discord to outline the actions it took during the Report Period to ensure that ineligible or illegitimate users are not able to access these Student Hubs.

Discord stated that when a ‘Student Hub is made available, Discord engages in a manual search to identify the official email domains associated with school email addresses.’ Discord stated that ‘only users who are able to provide, and verify access to, a school-issued email address may join a Student Hub.’

11. Question about information provided in NCMEC reports

In response to a question about the information Discord includes in its reports to NCMEC, Discord provided a list of information that it includes in a report to NCMEC. eSafety has correlated this with the types of information NCMEC defines as ‘actionable’ for law enforcement.

12. Question about implementing end-to-end encryption on Discord

eSafety asked Discord if it had carried out any work to implement end-to-end encryption during the Report Period. Discord stated that it had not done so for any part of its service.

13. Questions about Discord recommender systems

Design objectives of 'Server Discovery'

Discord was asked to describe the design objectives of the system it uses to organise the servers displayed on its 'Server Discovery' page. Discord reported that it didn't create specific objectives related to this surface but did provide five objectives that it explained combined 'both implicit and explicit goals' being:

- 'Increasing user engagement - have active users interact with the platform more.
- Increase user retention - have users be more likely to continue to use the service.
- User resurrection - increase the ease for an infrequent user to come back to the service.
- Improve new user signal-to-noise ratio - provide new users with high quality communities.
- Improve overall user experience of Discord.'

Primary signals used to recommend content to users

Discord was asked to list and describe the primary signals it uses to determine what content should be recommended to a user and to provide details of their relative importance in meeting the objectives it described in the previous answer.

Discord provided a list of signals in relation to 'Server Home', the surface that aggregates content from servers that users have not actively participated in recently, which included:

- overall age of server
- overall age of message
- number of user reactions to a message
- whether a message contains @everyone

eSafety has decided not to publish the full response provided by Discord due to confidentiality concerns.

When asked to explain the relative importance of these signals, Discord did not initially provide this information in response to the Notice. After a follow-up question from eSafety seeking the information required by the Notice, Discord said it was ‘unable to provide the relative importance of the primary signals used in the recommender system, because, by design, the recommender system is a machine learning model that adjusts the weights of factors based on outcomes.’

14. Questions about testing and updating its system to avoid amplification of harmful communication

Discord was asked questions about the measures it employs to test and update its system to improve the safety of its service and avoid amplifying harmful communities. Discord stated that its employees test ‘nearly every new Discord feature’ in advance of broader development. Discord also responded that it tests ‘Nearly every Discord feature or experiment’ before launching to the whole user base. Discord highlighted that it tests ‘heuristic recommendations’ twice a quarter, primarily through data examinations, and that it performs other testing, generally for new or updated features on an ‘as-needed basis’.

Discord also described a series of ‘key safety’ checks it performs on its recommender systems. On Server Home, Discord reported that it excludes messages from bots and messages posted from ‘age-restricted channels’. On ‘Top Message Push and Top Email Notifications’ Discord highlighted a range of checks.

15. Question about age-appropriate restrictions to Discovery recommendations

Discord was asked to describe the criteria it uses to evaluate communities that are harmful to children but do not violate community guidelines and any approaches it takes to not recommend, or recommend less often, these servers in ‘Discover’. Discord responded that servers that facilitate exchange of non-violative adult material must have age-restricted labels applied to applicable channels. These labels require users to confirm they are over legal age before viewing any content.

Discord stated that such servers are ineligible for 'Discover', 'making them less likely to be accessed by children through recommender systems.' Discord also reported that for a server to apply for Discovery, it must be a 'Community' server which means safety tools must be enabled. These safety tools include explicit media scanning and member email verification.

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

Discord responded that it does have internal metrics in place to assess the efficacy of its interventions to detect CSEA material including:

- Daily tracking of its automated CSAM detection systems including - the precision of the models; false positives; tasks the automated systems generate; outcomes of the automated systems; and proactivity rate of the systems.
- Using the above information to 'gain insight into [its] current state of operation', 'inform future improvements to its detection models and safety tools' and 'review the efficacy of its interventions and existing tools.'

17. Question about appeals against CSEA-related moderation

Discord was asked to provide the number of appeals made by users that had been banned or had content removed for CSEA-related violations. Discord stated it that it had received 483 appeals from Australian users during the Report Period and that 43 of those appeals were successful.

Discord noted that it faces 'data limitations in relation to data deleted because of retention periods'. It added that when a user is banned and their account deleted, the e-mail they appeal from will no longer be tied to a Discord user account – meaning that it will not have registration or session data for a user and cannot identify their location.

18. Questions about Safety by Design training for staff

Discord stated that it provides training on Safety by Design principles to its staff. Discord reported that its Trust and Safety team members attend 'extensive, live training' on moderating harmful content including CSAM and grooming. Discord added that each analyst

attends at least 50 hours of training, including live training and shadowing experienced reviewers, on topics such as:

- Determining if a user is underage
- Investigating potential grooming or child exploitation
- Reporting imminent threats to authorities for intervention
- Reporting CSAM to NCMEC.

19. Additional information

In response to an opportunity to provide further information and context to any of its responses to the questions asked in the Notice, Discord provided an overview of how it operates compared to other platforms stating that it ‘is a voice, video, and text communication platform used by over 150 million monthly active users across the globe to hang out and talk with their friends and communities.’ Discord explained that its users tend to use the service to connect with friend groups and shared interest communities unlike other platforms that are used to communicate with mass audiences. Discord also reported that its content moderation approach is based on three pillars:

1. ‘User level’ – empowering users to control their communities using the user controls and safety settings available to them ‘including the ability to restrict who can send a user a direct message, who can add a user as a friend, what servers a user joins, the ability to block specific users, the ability to automatically scan and delete direct messages with explicit images, the ability to restrict and control who joins a user’s private server, the ability to control who has the ability to invite individuals to a user’s server, the ability to have server invites automatically expire after a period of time or number of uses, the ability to restrict visibility of channels to other individuals who join a user’s server, and the ability to choose security and verification levels for a user’s server.’
2. ‘Community level’ – empowering volunteer administrators and moderators to set the tone for their communities and ensure the safety and wellbeing of their communities.
3. ‘Platform level’ – ‘Discord’s Community Guidelines and policies provide the foundation for keeping the platform safe’ and apply to all users, every action and all content.

Discord added that it ‘integrates with the Crisis Text Line – a nonprofit that provides 24/7 text-based mental health support and crisis intervention via trained volunteer crisis counselors.’ When a user reports self-harm within the Discord mobile app they are automatically provided with information on how to contact Crisis Text Line.

Discord also highlighted that it collaborates with the Digital Wellness Lab at Boston’s Children Hospital to ensure the latest scientific research and best practices are incorporated into Discord’s approach to teen safety and belonging.

Discord added that its Moderator Academy also ‘provides instruction to moderators on topics pertaining to teen safety and dealing with sensitive and triggering topics.’



[eSafety.gov.au](https://www.esafety.gov.au)