Child Sexual Abuse Material (CSAM)

Identification and Reporting for U.S. Based Companies

January 2022

TECH COALITION

CSAM Identification & Reporting for U.S. Based Companies

Child Sexual Abuse Material (CSAM) reporting is a complex global issue governed by multiple unique jurisdictional definitions and requirements. The Tech Coalition produced this paper to provide a clear overview of current CSAM identification efforts and the reporting regime for US-based companies. This paper is for educational purposes only and focused exclusively on current practice. We hope that you find it helpful.

Although there is no single legal definition for CSAM, this term generally refers to sexually explicit imagery involving a child[1]. CSAM includes still images, videos, and illustrated, computer-generated or other forms of realistic depictions as well as live streaming broadcasts of a human child in a sexually explicit context, or engaging in sexually explicit acts. It also includes links to third-party sites that host child sexual exploitation material. It may also include digital or computer generated images indistinguishable from an actual minor.

Currently, global CSAM reporting is led by US-based Electronic Service Providers (ESPs) reporting to the US-based, private, non-profit organization, the National Center for Missing and Exploited Children (NCMEC). Although US law broadly shields ESPs from civil and criminal liability for user generated content, there is a significant exception to this immunity for ESPs with "actual knowledge" of CSAM. In 2008, the US Congress passed the PROTECT Our Children Act[2] to create a reporting requirement for tech companies with "actual knowledge" of facts and circumstances of child exploitation on their services. This same law designates NCMEC's Cybertipline as the clearinghouse for these Cybertip reports. The ESPs report to NCMEC and NCMEC makes the reports available to the relevant US or Foreign law enforcement agency for their independent review and possible further action.

In 2021, NCMEC's CyberTipline received over 29 million CSAM reports. Though the majority of CyberTipline reports are received from US-based providers, NCMEC also receives reports from non-US-based ESPs. Over 90% of CSAM reports submitted to NCMEC's CyberTipline in 2021 concerned incidents that involved an individual located outside of the United States. NCMEC makes reports originating outside the US available to more than 140 countries and territories either directly to the national police force in country or via US-based federal law enforcement agencies who may forward reports internationally. A full country by country breakdown of 2020 report origination can be found here.

ESPs operate a number of different types of services and platforms that may be impacted by CSAM, from consumer cloud storage to messaging, and social media to video chat. Each service is unique and ESPs have varying levels of capacity to combat CSAM based on a range of factors, including their size and maturity. No two ESPs operate the same with respect to CSAM. There are, however, as outlined below, broad similarities in the technical approach that are helpful to understand.

A. How do ESPs detect CSAM?

Regardless of jurisdiction, ESP's may detect CSAM on their platforms through three primary, voluntary mechanisms:

1. Hash-Based Detection for "Known" CSAM

Current technology allows ESPs and NCMEC to create and assign unique numerical "hashes" or digital fingerprints to images that are confirmed as CSAM. ESPs may then use that same technology to automatically screen uploaded imagery CSAM using databases of these hashes. This technology generates the vast majority of CSAM identification and reports. Facebook has stated that more than 90% of its reports to NCMEC between October and November 2020 concerned shares or reshares of previously detected content.

One example of this hash-based detection technology is Microsoft's PhotoDNA that Microsoft developed in cooperation with Dartmouth College in 2009. Google developed CSAI Match, a video hash-based detection tool, in 2014. In 2019, Meta developed PDQ and TMK+PDQF, two open-source photo and video-matching technologies that detect identical and nearly identical photos and videos. To accelerate industry wide efforts to stop CSAM, Microsoft, Google and Meta make all of these tools available to other industry members and qualified nonprofits free of charge.

Hash-based detection works only as well as the quality and accuracy of the underlying hash-sets. Industry currently relies upon both their own self-created hash sets based on imagery identified by their systems and hash repositories created by several different NGOs, including NCMEC, Thorn, the Internet Watch Forum, and the Canadian Center for Child Protection. ESP hashing tools assign hashes to user generated imagery on their platforms, compare those hashes against these hash-sets of known CSAM, identify any matches, then flag that material for further review or immediate take down and reporting.

Hash-based detection is used for both still and video images. Hash technology for still images is mature and standardized. There is, however, no global industry standard for video hashes at this time. Current video CSAM detection tools use different hashing protocols that are not interoperable. This means that ESPs are handicapped in their efforts to identify and report video CSAM because, unlike still image detection, their

video detection tools cannot use a common database of known video CSAM hashes. The Technology Coalition and its members are currently working with NCMEC and Thorn to create a video hash "translator" that will allow ESPs to share hashes so that known video CSAM may be more quickly identified and reported.

Hash-based detection only detects images that have been previously identified as CSAM and hashed. Nevertheless, this technology accounts for the vast majority of CSAM identification and significantly reduces the revictimization of survivors of child sexual abuse resulting from the republishing of known CSAM.

2. User or Third Party Reporting

In addition to hash-based detection, ESPs may provide reporting mechanisms for users and third parties to report CSAM on their platforms. This reporting may cover previously identified CSAM that has already been hashed or newly published CSAM that has not previously been identified. Although there is tremendous variation in the substance and accuracy of these reports, ESPs continue to take steps to encourage user reporting of potentially abusive content on their platforms.

3. Machine Learning Classifiers

As discussed previously, **hash-based detection is only effective for previously identified CSAM**. For new, not previously identified CSAM, ESPs may deploy machine learning classifiers. These classifiers flag suspected CSAM, which is then confirmed by specialist human review. If this not-previously-identified material is confirmed to be CSAM, it is assigned a hash that can then be used by hash-based automatic detection systems to prevent the further dissemination of the new material.

This type of technology requires significant resources and data to develop and test. Google makes its machine learning classifiers available to industry through the Content Safety API. THORN also includes a classifier to identify previously-undetected CSAM in their Safer detection service. This helps ESPs identify content likely to contain CSAM faster and prioritize it for human review.

4. Human Review

There is significant variation among ESPs with respect to the use of human moderators in the detection and reporting of CSAM. For hash-based detection, reporting may be automated or sent to human moderators for review, classification and reporting. With machine learning classifiers, flagged imagery goes through human review to confirm that it is indeed previously-undetected CSAM. Similarly, when a user or third-party reports

alleged CSAM directly to an ESP, those reports go to human moderators for review and confirmation.

Human review is a costly and time-consuming aspect of CSAM. Perhaps the most significant cost of human review is the impact on the individual specialists conducting the review. ESPs continue to take steps to reduce the negative impacts on moderation staff through technology and wellness initiatives.

B. What do ESPs do when they detect CSAM?

Each ESP has unique user agreements and content policies that govern their internal procedures for handling confirmed CSAM. Generally, once an ESP confirms that an uploaded image is CSAM, then it will take the following steps:

1. Remove access to the CSAM if it has been published.
2. Report the image to NCMEC's CyberTipline.
3. Preserve the suspected CSAM in a secure location with limited access for 90 days.[3]

ESPs may also take one or more of these additional steps:

1. Immediately and permanently suspend the offending user's account.
2. Prohibit the offending user from creating any new accounts in the future.

C. How/What Do ESPs report to NCMEC's CyberTipline?

In the United States, the PROTECT Our Children Act[4] requires ESPs to report suspected CSAM to NCMEC's CyberTipline. The law enables the voluntary reporting of the following categories of information to be included as part of the 'facts or circumstances' of the CyberTipline report: information about the involved individual responsible for the apparent violation, including identifying information (e.g., email address), and location information (e.g., IP address), how the provider became aware of the violation, the visual depiction of child sexual abuse, and the complete communication containing the visual depiction.

The ESP reporting process to the CyberTipline may either be done manually or through an online interface that NCMEC provides to registered ESPs. ESPs may register with the CyberTipline to access a secure reporting webform or API provided by NCMEC.

In 2020, Tech Coalition member companies provided 98% of all reports to the NCMEC CyberTipline. (TC Annual Report).

D. How does NCMEC's Cybertipline process the reports?

In March 1998, NCMEC created the CyberTipline to serve as an online mechanism for members of the public and electronic service providers (ESPs) to report incidents of suspected child sexual exploitation. The CyberTipline is not limited to CSAM reporting. In addition to CSAM, the CyberTipline receives reports of child sex trafficking, online enticement of children for sexual acts; extra-familial child sexual molestation; child sex tourism; unsolicited obscene materials sent to children; misleading domain names; and misleading words or digital images.

In 2021, NCMEC received an average of 80,000 CyberTipline reports a day, totaling over 29.3 million for the year. In addition to the US-based ESPs that are required to report to the CyberTipline by US law, approximately 50 EU-based companies have voluntarily registered to report suspected CSAM to the CyberTipline.

One of NCMEC's central goals as a clearinghouse of CyberTipline reports is to determine the potential location of a reported incident so the report can be made available to the appropriate law enforcement in the United States or globally. As part of this review last year, over 90% of CyberTipline reports were made available to international law enforcement for review. Currently, NCMEC has secure, encrypted connections with law enforcement in over 140 countries and territories, including each country in the EU, as well as Europol, to make reports available around the world.

NCMEC also works closely with Interpol to make elements of CyberTipline reports available in countries where it does not have a local law enforcement connection. Due to the volume of international reports, NCMEC staff are not able to conduct individual review of reports that geo-locate to a non-U.S. location, except in a handful of exceptions. Instead, these reports are auto-referred to an appropriate international law enforcement agency based on IP address, phone number, or location information provided within the report.

For the approximately 5% of CyberTipline reports that the CyberTipline makes available to U.S.-based law enforcement, NCMEC has staffing capacity to conduct hands-on review and analysis that may include one or more of the following:

1. Assessing the immediate risk to a child;
2. Determining if the suspect, victim, user name, or other unique identifiers contained within the report have been previously reported to the CyberTipline;
3. Determining the geographic location for the CyberTipline report; and
4. Depending on staffing capacity and information provided by the ESP, reviewing one or more of the uploaded files submitted in the CyberTipline report.

Additionally, NCMEC staff may conduct open-source queries regarding the reported information. As a part of the CyberTipline triage process, certain reports are escalated based on information provided by the ESP, while other reports may be provided to law enforcement in an informational capacity.

E. How Does NCMEC's CyberTipline Generate and Maintain CSAM Hash Databases?

NCMEC's CyberTipline contains a repository of all CSAM reported to it **since 1998**. This serves as the basis for NCMEC's hash lists, which are used internally to classify incoming content and also shared externally with ESPs through multiple different hash-sharing platforms. Currently, **NCMEC shares over 5 million hashes of CSAM and approximately 244,000 hashes of sexually exploitative content with ESPs** through these platforms.

NCMEC endeavors to supply hashes in multiple different hash types to accommodate differing systems used by ESPs. This is time-consuming and costly to maintain given the expenses involved in generating different hash types for a single file, the necessity for direct access to volumes of CSAM material, and the need for the CyberTipline to determine at the outset which hash types to support.

F. What Resources Are Required to Operate NCMEC's CyberTipline?

The CyberTipline incorporates various external comparison tools to enhance image matching and file tagging including: Microsoft's PhotoDNA, Videntifier video matching service, Google Content Safety API, and Thorn's Safer product. NCMEC is constantly evaluating new technology and upgrades to ensure its image matching and file tagging procedures are incorporating the most advanced tools available on the market.

Given the immense volume and technical complexity of reports being submitted to NCMEC's CyberTipline, substantial resources are required to run the CyberTipline operations. Currently there are 117 employees working in NCMEC's Exploited Children Division (ECD). In addition to generous technology contributions that NCMEC receives each year to support its ECD work, NCMEC expends approximately $20 million a year to operate the CyberTipline.

Conclusion

As previously stated, the Tech Coalition produced this paper to provide a clear explanation of current methods to detect CSAM and the reporting regime for US-based companies. It is for educational purposes only and focused exclusively on current practice. For additional information on CSAM reporting, please refer to this summary of the Tech Coalition's March 2021 summit on The Next Frontier of Reporting. You can also find additional resources at THORN, WeProtect, End Violence and NCMEC.

Footnotes:

[1] In the Temporary Derogation to Directive 2002/58/EC, the European Parliament has defined CSAM to encompass "Child Pornography" as defined in

Article 2(c) of Directive 2011/93/EU
(c)'child pornography' means:
(i) any material that visually depicts a child/ engaged in real or simulated sexually explicit conduct;
(ii) any depiction of the sexual organs of a child for primarily sexual purposes;
(iii) any material that visually depicts any person appearing to be a child engaged in real or simulated sexually explicit conduct or any depiction of the sexual organs of any person appearing to be a child, for primarily sexual purposes; or
(iv) realistic images of a child engaged in sexually explicit conduct or realistic images of the sexual organs of a child, for primarily sexual purposes;

And "pornographic performance" as defined in Article 2(e) of Directive 2011/93/EU
(e) 'pornographic performance' means a live exhibition aimed at an audience, including by means of information and communication technology, of:
(i) a child engaged in real or simulated sexually explicit conduct; or
(ii) the sexual organs of a child for primarily sexual purposes;

In the United States, Title 18 United States Code § 2256(8) defines child pornography (CSAM) as "any visual depiction, including any photograph, film, video, picture, or computer or computer-generated image or picture, whether made or produced by electronic, mechanical, or other means, of sexually explicit conduct, where—
(A) the production of such visual depiction involves the use of a minor [person under 18] engaging in sexually explicit conduct;
(B) such visual depiction is a digital image, computer image, or computer-generated image that is, or is indistinguishable from, that of a minor engaging in sexually explicit conduct; or
(C) such visual depiction has been created, adapted, or modified to appear that an identifiable minor is engaging in sexually explicit conduct.

[2] 18 U.S. Code § 2258A
[3] 18 U.S. Code § 2258A
[4] 18 U.S. Code § 2258A