The Future of Trust & Safety: Child Safety Investigations

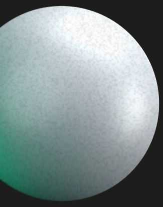
FALKOR

Table of Contents

Introduction

Let's start with a fact that's often overlooked when talking about Child Safety: it's not just about CSAM. Its definition also encompasses such issues as Grooming, Sextortion, Self-Generated Content, Deepfakes, Non-Consensual Intimate Imagery (NCII), Child Sex Trafficking, Child Labor Trafficking, Child Marriage, and many many more. It’s quite a lot to consider and address.

Nevertheless, as a professional working in the Child Safety field or handling related escalations, please always remember: you are not alone in this battle.



The importance of investigating Child Safety issues

Child Safety investigation is necessary for two main (and quite obvious) reasons:

1. Risk to a platform's reputation

If a platform can’t tackle this problem in a timely manner and it gets known (to the press or Law Enforcement, for instance), it might be detrimental. For small and emerging apps, it might be hard to recover from bad publicity, maintain a positive brand image for current and future users, and change a circulating narrative.

2. Negative impact on user retention

Partially related to the issue described above, the experience of seeing Child Sexual Abuse Material or being groomed (for underage users) is shocking, harmful, and in no way pleasant; with the vast array of social media apps, the users may choose to go elsewhere, where they will be feeling safer.

Trends observed in 2022

1. Generative AI (yes, we all heard about the hype around its use - but let's consider its "dark" side):

One of the now well-known (mis)uses of Generative AI (or GenAI) is the creation and distribution of "fake" CSAM imagery. GenAI is used to generate a child's body, while the real victim's face is attached to it. It can reach an enormous scale by using the same face for multiple images that circulate online. A child, who by now may be an adult, will be victimized hundreds, if not thousands, of times.

Additionally, GenAI has been (mis)used to engage in grooming conversations with minors. How? It can create a script that sounds appropriate for a child. It gives the impression that the young user is communicating with someone their own age. Both machine and human moderators will think the same. In reality, it can easily be an older person posing as a 13-year-old. Previous to this change, certain patterns made it easier to detect an adult pretending to be a minor. It will now be much more difficult.

2. CAP sites:

In 2022, INHOPE has observed the emergence of Child Abuse Pyramid (CAP) Sites. This is a new type of commercial site that uses a particular form of "invitation" system to access CSAM content. How does it work?

Imagine a referral system in some companies, "refer a friend and get a bonus". Or in computer games - earn points to open new levels. Or convert your frequent flyer miles to enter the business lounge. All these examples are harmless.

In CAP sites the idea is the same but the outcome is different. Users are encouraged to share their personal links to invite others to the CAP site. The more invitations they share, the more "points" they accumulate to access CSAM content. It’s quite concerning, as this “reward” system incentivizes bad actors to expand their network by adding a game element in it, so they can find and bring new CSAM consumers faster.

Child Abuse Pyramid (CAP) Sites is a new type of commercial site that uses a particular form of "invitation" system to access CSAM content.

3. Livestreaming:

The name speaks for itself; however, let's go deeper into the now-known categories. As described in a comprehensive report by NetClean and later summarized by WeProtect Global Alliance, the abuse through livestreaming may take various forms. Excluding the voluntarily self-produced live-streamed material, we wanted to bring your attention to two categories:

1- Child sexual exploitation and abuse is happening offline and streamed in real-time to remote viewers.

2 - One or more children have to "perform" sexual acts in front of a camera, while someone is "orchestrating" it by either typing the instructions for the moves or dictating it from behind the camera.

Unlike self-generated material, in these cases one or several individuals are financially motivated to facilitate the abuse, and a person on the receiving end is willing to pay to watch the "performance" live. A minor may be forced or coerced into doing it; in some cases, the victims don't know that they are being recorded.

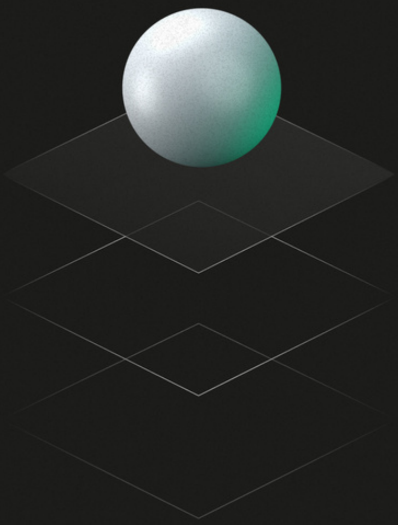
It's important to emphasize that this type of abuse is driven by economic inequality and Covid-19 played a role in its increase. According to the statistics released by WeProtect Global Alliance, "a 265% increase in livestreaming was recorded in the Philippines during the quarantine period."

The main problem with livestreaming for T&S professionals is that its detection is extremely difficult. Since the abuse happens in real time, usually no evidence is left afterward. In addition, it's rare for the viewers to report it, as they pay to be part of the "performance".

Questions for SMEs and T&S professionals to consider

The trends for each online platform are different depending on its type, audience age and geography, available features, and main focus. However, the importance of addressing Child Safety-related issues in a timely manner is equally high for everyone. To ensure your platform is tackling its internal threats, ask yourself:

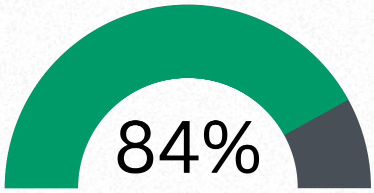
- What are our platform-specific trends when it comes to Child Safety?
- What is the Average Handle Time (AHT) between content detection, moderation, and escalation to Law Enforcement?
- Is our platform equipped to prevent bad actors from returning after they were permanently banned?



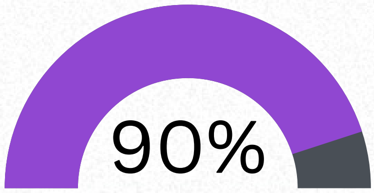
Learn how you can supercharge your Child Safety investigations with Falkor

Statistics for 2022

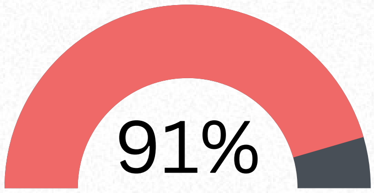
According to INHOPE's Annual Report, 84% of CSAM reports were of never-before-seen material, depicting new child victims of sexual abuse. The alarming fact is the age of the victims. 1% of the images were of infants (0-2 years old), with children as young as 4 months old. 9/10 victimized children were between 3-13 years old. Females are still a predominant majority, 91%.



Never-before-seen material

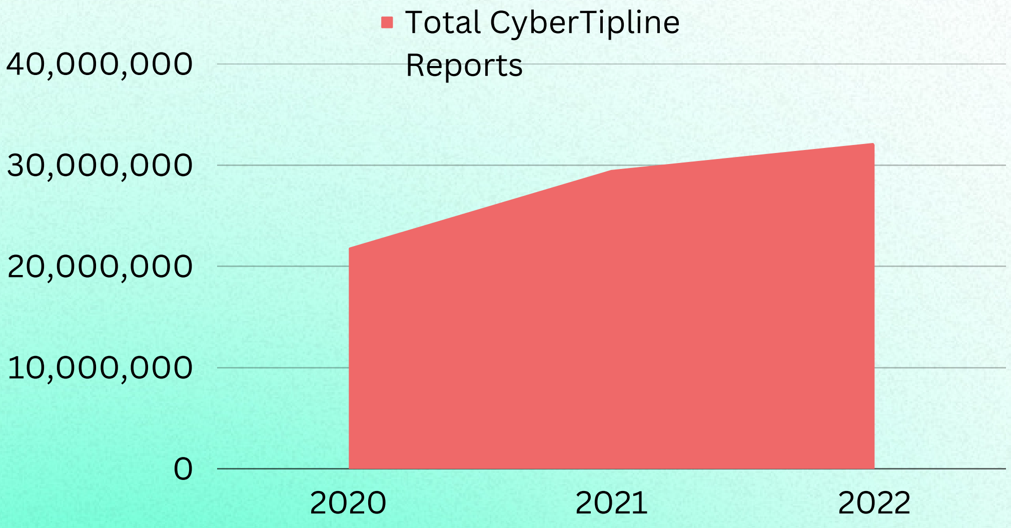


Victims aged 3 to 13



Female victims

Increase in CyberTipline Reports in the US:



Statistics for 2022

In addition to these numbers, NCMEC's Annual Report also shows an increase in CyberTipline reports in comparison with 2020. However, what we want to bring your attention to is a discrepancy between actionable and informational reports compared to their total number.

For instance, Internet Crimes Against Children Units received 892,370 reports in 2022, and only 491,655 were acted upon. For law enforcement, out of 1,465 reports received during that year only 3 were considered informational. Overall, only 50% of the reported cases were marked as actionable and even less qualified as informational. On a global scale, it means that it takes longer to find actionable reports, prevent incidents, and help rescue a victim on time.

Law Enforcement Recipients of CyberTipline Reports	2022
Internet Crimes Against Children (ICACs) Units	Actionable: 491,655 Informational: 400,715 Total: 892,370
Local Law Enforcement	Actionable: 1,462 Informational: 3 Total: 1,465
Federal Law Enforcement	Actionable: 1,356,988 Informational: 997,475 Total: 2,354,463
International Law Enforcement	Actionable: 13,995,567 Informational: 15,010,413 Total: 29,005,980

New and upcoming regulations

Currently, Child Safety is one of the top priorities for all User Safety regulations. Aside from the well-known and discussed Digital Services Act (DSA) in Europe and Online Safety Bill in the UK, Australia already has its own Online Safety Act.

Those platforms with a user base in the US and US-based platforms - brace yourselves. In contrast to other countries, the US doesn't yet have a centralized regulatory system - most laws are state-specific, meaning each state can set up its own regulations and preventive measures regarding Child Safety. We recommend looking out for the Children's Online Privacy Protection Act (COPPA), California Age-Appropriate Design Code Act (the Act), and 18 U.S. Code § 2258 for all states.

While we won't debate the regulations, the main point we wanted to emphasize is this: platforms that don't moderate or report Child Safety cases will soon pay a high price. Aside from the potential financial and reputational damage to the company, let's not forget about the children who will be repeatedly traumatized so long as their images circulate. As professionals and as a society, we owe it to ourselves and to victims to stop abuse.

Recommendations to avoid potential penalties

Assess the risks associated with current and new features - consider how regular users and bad actors might misuse and abuse them.

Re-evaluate Community Guidelines and re-enforce some policies - if minors' safety is prioritized, make sure the rules reflect that priority.

Ensure compliance with the laws of countries and/or states where the user communities originate - in this case, ignorance is not bliss.

Questions for SMEs and T&S professionals to consider

As NCMEC's findings have shown, 50% of the reports don't provide sufficient information for further investigation. With the new and upcoming regulations, we expect an increase in Child Safety-related escalations. To ensure your platform is compliant and prepared, ask yourself:

→ What are the loopholes and gray areas in our guidelines and features? Is it easy for bad actors to bypass our AI detection?

→ If our platform is for adults only, how do we ensure there are no underage users? If there are any, do we have an enforcement plan in place to detect and prevent potential Grooming and other issues?

→ What is the quality of the current Law Enforcement reports we share? What can be improved?

Learn how Falkor can supercharge your reports.

The current state of T&S investigations

When moderators encounter suspicious content, do T&S teams investigate it?

Somewhat; but going slightly deeper will depend on the team size, skills, and available resources. Conducting even basic open-source investigations is not that widely common.

When a Child Safety violation is detected and reported on a platform, what happens next?

In a nutshell, nothing. As part of the platform's responsibility, the team notifies law enforcement of the problematic user and content.

The current approach does not solve the end problem. Until they are detected and reported, the bad actors will continue to violate the platform's policies. It is possible that they will use existing accounts or create new ones, bypass ID verification (if there is one), spread CSAM material, and prey on minors - business as usual.

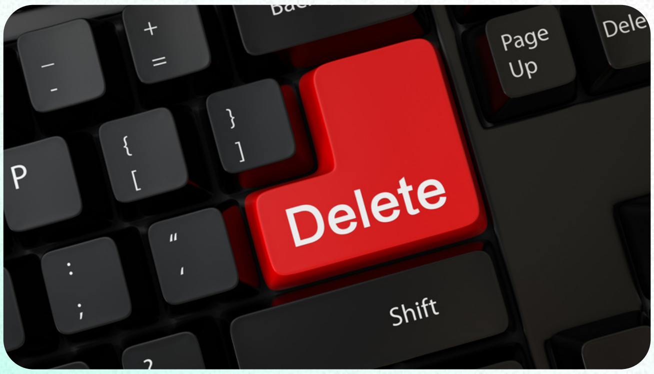
We will discuss a high-level workflow for Child Safety cases - how many small and medium-sized teams handle them.

In summary, the workflow consists of three main steps: detect, remove, and report.

1. Content detection

Content that may be harmful to users can be detected in a few ways, depending on the platform's features. It can be done with the help of AI, based on User Reports, or via manual detection. Either way, the content will be reviewed according to internal policies to determine if a violation exists. Let's say a child safety violation was discovered.

2. Content removal



The next step would be to remove the content and either suspend or permanently ban this user. A moderator's decision will be affected by the platform's internal policies - in some cases, they will need approval from their managers first. Whenever possible, harmful content should be removed immediately so that fewer users are exposed to it.

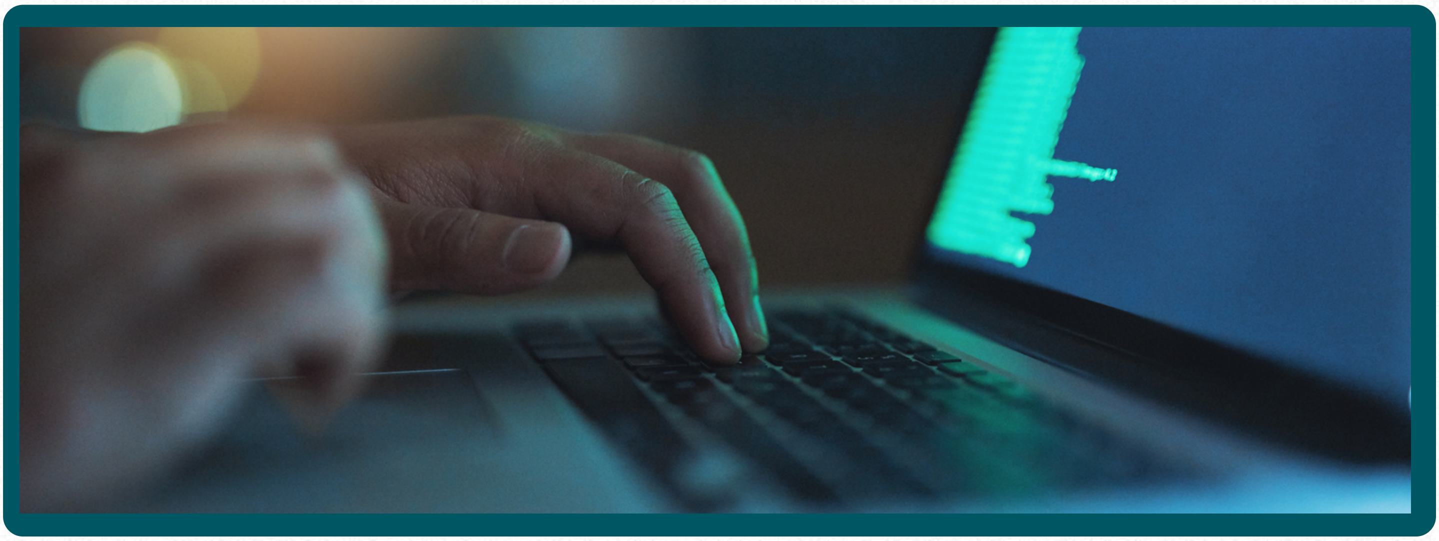
Restricting a profile enables the team to extend its investigation and determine whether it constitutes a Child Safety violation. Furthermore, that same user can no longer distribute this type of content on the platform in the future. Let's assume that the content was confirmed as CSAM.

3. Content reporting

If a decision maker in the company confirms that the content is positive for CSAM, the user will be permanently banned from the platform and reported to the relevant law enforcement agency (e.g. NCMEC in the USA). The person or team that handles Law Enforcement escalations must gather the necessary "evidence" and fill out the online reporting form within a certain period of time.

However, as we can see in the 2022 statistics, Law Enforcement specialists drown under the volume of reports they receive. In theory, it's not the platform's concern. Nevertheless, if we want to combat Child Safety on all fronts, let's work together.

Investigations can be conducted in-house and on a low budget. Your team can quickly acquire the needed OSINT (open-source intelligence) skills, along with an understanding of what's critical to report to law enforcement agencies so they can tackle real-life threats. That is why after removing violating content from your platform, the ideal “extra” step is to investigate it.



What can be done to get rid of the bad actors altogether?

Find the root cause of the problem

- Investigating bad actors - Where these accounts are coming from? Can we trace their origin via static IP?

- Network analysis - Are they connected, or is it one person behind? What are the patterns they have in common?

- Proactive research - Did they reach out to other users? How many users might be potentially affected?

These are some questions that your company can use for guidance. It's not that complicated. And the best part: you already have all the data you need. Utilizing, analyzing, and enriching it correctly is the missing piece.

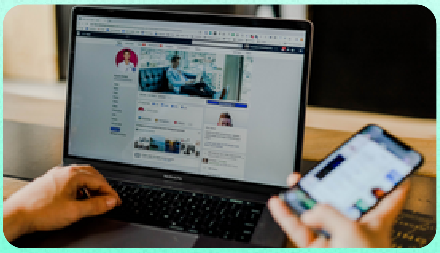
Discover the future with Falkor

Visit falkor.ai or contact us at hello@falkor.ai

More resources:

From content moderation to proactive investigations: Empowering platforms for a safer online environment

See no evil, hear no evil: siloed trust and safety teams



Eyes wide shut - investigation challenges for trust and safety teams