Terre des Hommes

DETECTION OF CHILD SEXUAL ABUSE MATERIAL

Why do we need detection?

Unless detected, Child Sexual Abuse Materials (CSAM) continue to be disseminated, leaving victims in an **ongoing trauma** and fear to be recognised, in **violation of their privacy**. Without detection, the problem will continue to grow, as CSAM is found by and shared between offenders. The huge **volume** of currently existing CSAM requires the use of **technological tools** to detect and remove them. Those tools, such as PhotoDNA, have been deployed for over a decade and are proven to work and are set up to search for matches against an established list of known CSAM.

Terminology explained

Known CSAM = material confirmed to constitute child sexual abuse material

New/Unknown CSAM = material that constitutes child sexual abuse material, but not (yet) confirmed as such by an authority

Current reporting mechanisms are not enough

Public reporting will never be sufficient to identify CSAM due to several barriers to reporting, including:

- the fact that many victims are prepubescent and, therefore, **too young** to report or are **threatened** by offenders not to report
- **Shame** and taboo in encountering such material prevent reporting

What about preventative measures?

Although awareness and education efforts contribute to increasing public reporting, it will **not resolve all the underreporting issues.**

THE SCALE OF THE ISSUE REQUIRES TECHNOLOGY

How big is the problem?

In 2022, **88.3 million files** containing CSAM were reported to the NCMEC Cybertip line. **This represents just the tip of the iceberg,** since not all companies report due to lack of regulation.

Proactively searching for CSAM is vital.

Proactive detection of CSAM leads to a **substantially higher volume** of identified and removed CSAM. For example, the Canadian Project Arachnid's automated proactive web crawling detection tool of known CSAM or close matches processed 160 billion+ images between 2017 and September 2023. Proactive search is able to identify CSAM at a scale that meets the **volume of CSAM** in circulation.

It is equally **important to detect previously unknown CSAM**, in this way, new victims could be identified and rescued.

What is the impact of detecting CSAM?

Detecting CSAM reports from platforms is estimated to lead to over

800
arrests of suspected child sex offenders
(UK Gov)

1200
children safeguarded on average every month in the UK

Terre des Hommes

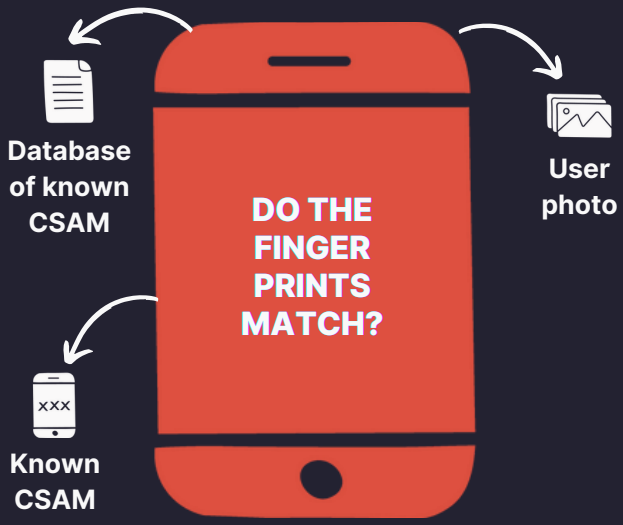
TOOLS AT OUR DISPOSAL

PHOTODNA

PhotoDNA technology creates a unique digital fingerprint (known as a "hash") of an image. This hash is then compared against fingerprints (hashes) with images in a known database of CSAM to find potential matches. This is not facial recognition software and **cannot be used to identify a person or object** in an image. A hash is not reversible, and therefore cannot be used to recreate an image.

Example of using PhotoDNA technology in a chat app:

A server scans the fingerprint (hash) assigned to a user's image against a database of known CSAM with hashes to determine if it is a copy of a hash in that database. Here, scanning occurs at the stage where a user attempts to upload a photo from their device gallery to a chat interface. Other scanning techniques are already being used by chat apps to scan for malicious messages and viruses.



Browse and upload image to chat from device gallery

CLASSIFIERS

Classifiers to detect unknown CSAM:

Machine learning AI classification models are also being used to detect unknown CSAM. By training these classifiers, or algorithms, these technologies are **trained to only identify, predict and distinguish if a piece of media constitutes CSAM or not.** This means it does not know what is in the picture, except whether it is likely CSAM or not. It can make the difference between criminal material and innocent pictures. Detection tools are built for a purpose and are hard and costly to re-purpose.

Behavioural classifiers:

An **algorithm** that identifies child users based on patterns users produce when clicking pictures. Children tend to have different behavioural patterns online e.g. choosing pictures that they are more exposed to. This **algorithm uses cognitive and behavioural classifiers** to differentiate between adults and children. Experiments show that this can be an effective tool to identify and protect children from exposure to inappropriate content or contacts that might put them at risk of grooming.

THE NUMBER ONE DEMAND FROM VICTIMS: REMOVING THEIR IMAGES FROM THE INTERNET.

Survivors of abuse and/or exploitation not only endure **trauma** from the abuse, but also face the **distressing knowledge that images or videos of their ordeal circulate online.** This can elicit various responses, like developing **PTSD**, from the survivors themselves but also from their families (C3P, NCMEC, RAINN).

Detecting images is protecting victims' privacy, to prevent further dissemination.

What about my privacy?

Detection technology is built for purpose and to minimise privacy invasion, with privacy by design as a guiding principle. Check out the Fact-Check on detection technology and many other tools like Microsoft Project Artemis, YouTube CSAI , Match, Google Content Safety API

Sources: Child Safety by Design 2022, European Commission Comments on IA 2023, Nezhad, M. M. and Mehrnezhad, M. (2018) 'A child recognition system based on image selection patterns'. Safer.