Expanded Protections for Children

August 2021

At Apple, our goal is to create technology that empowers people and enriches their lives — while helping them stay safe. We want to help protect children from predators who use communication tools to recruit and exploit them, and limit the spread of Child Sexual Abuse Material (CSAM).

Apple is introducing new child safety features in three areas, developed in collaboration with child safety experts. First, new communication tools will enable parents to play a more informed role in helping their children navigate communication online. The Messages app will use on-device machine learning to warn about sensitive content, while keeping private communications unreadable by Apple.

Next, iOS and iPadOS will use new applications of cryptography to help limit the spread of CSAM online, while designing for user privacy. CSAM detection will help Apple provide valuable information to law enforcement about collections of CSAM in iCloud Photos.

Finally, new additions to Siri and Search provide parents and children expanded information and help if they encounter unsafe situations. Siri and Search will also intervene when users try to search for CSAM-related topics.

These features are coming later this year, in updates to iOS 15, iPadOS 15, watchOS 8 and macOS Monterey.[1]

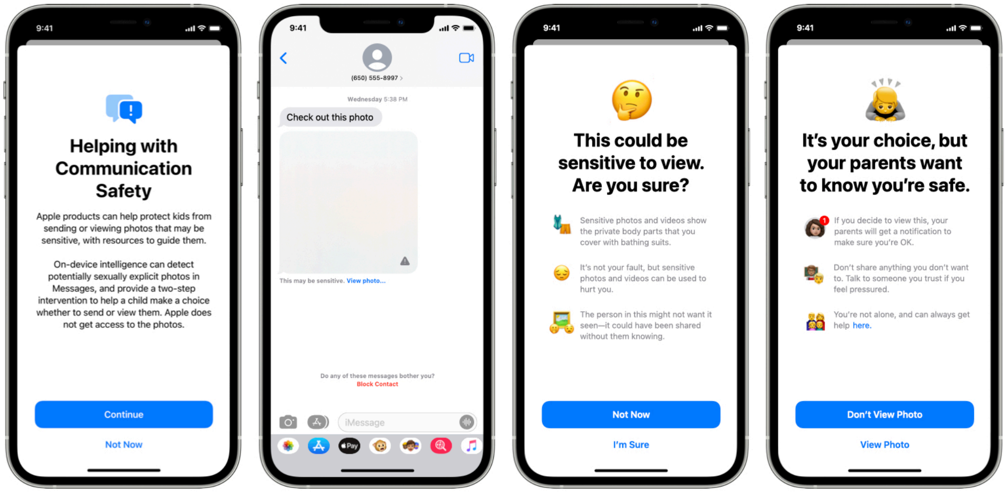
This program is ambitious, and protecting children is an important responsibility. Our efforts will evolve and expand over time.

[1] Features available in the U.S.

Communication safety in Messages

The Messages app will add new tools to warn children and their parents when receiving or sending sexually explicit photos.

When a child receives this type of content the photo will be blurred and the child will be warned, presented with helpful resources, and reassured it is okay if they do not want to view this photo. As an additional precaution the child can also be told that, to make sure they are safe, their parents will get a message if they do view it. Similar protections are available if a child attempts to send sexually explicit photos. The child will be warned before the photo is sent and the parents can receive a message if the child chooses to send it.



Helping with Communication Safety

Apple products can help protect kids from sending or viewing photos that may be sensitive, with resources to guide them.

On-device intelligence can detect potentially sexually explicit photos in Messages, and provide a two-step intervention to help a child make a choice whether to send or view them. Apple does not get access to the photos.

(650) 555-8997

Wednesday 5:38 PM

Check out this photo

This may be sensitive. View photo...

Do any of these messages bother you?
Block Contact

This could be sensitive to view. Are you sure?

Sensitive photos and videos show the private body parts that you cover with bathing suits.

It's not your fault, but sensitive photos and videos can be used to hurt you.

The person in this might not want it seen—it could have been shared without them knowing.

It's your choice, but your parents want to know you're safe.

If you decide to view this, your parents will get a notification to make sure you're OK.

Don't share anything you don't want to. Talk to someone you trust if you feel pressured.

You're not alone, and can always get help here.

This new feature helps warn children and their parents when sending or receiving sexually explicit images.

Messages uses on-device machine learning to analyze image attachments and determine if a photo is sexually explicit. The feature is designed so that Apple does not get access to the messages.

This feature is coming in an update later this year to accounts set up as families in iCloud for iOS 15, iPad OS15 and macOS Monterey.[1]

[1] Features available in the U.S.

CSAM detection

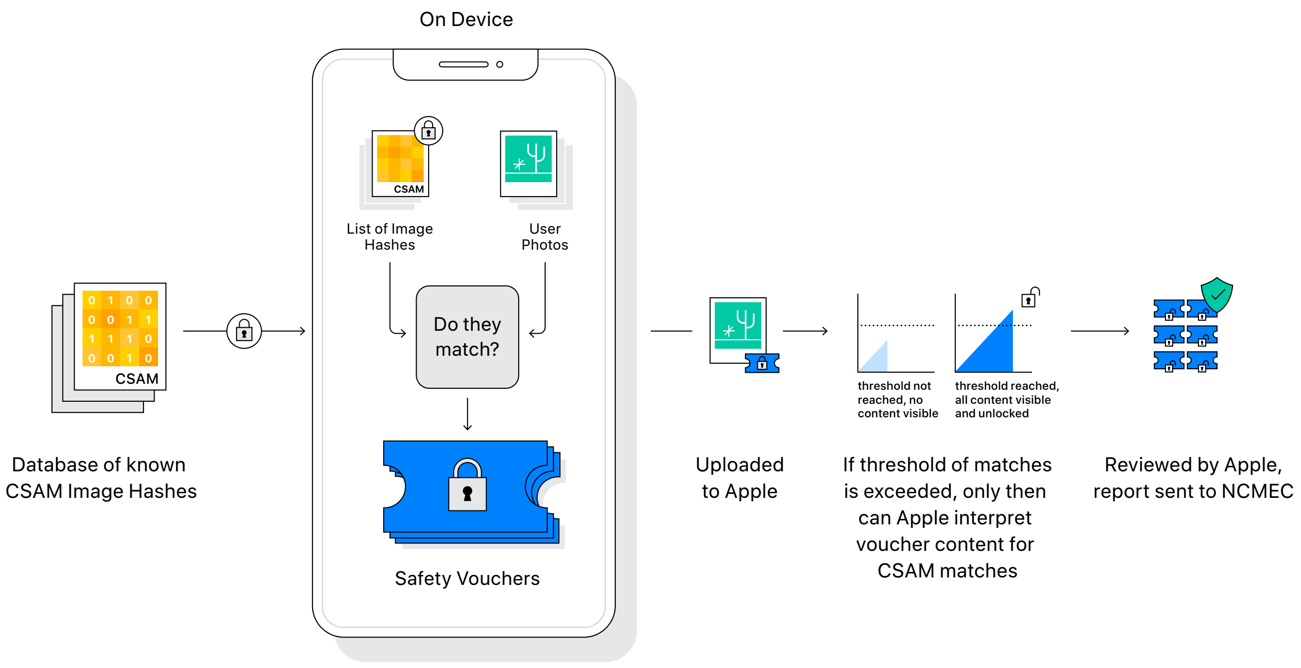
NATIONAL CENTER FOR MISSING & EXPLOITED CHILDREN

The National Center for Missing & Exploited Children is a private, non-profit 501(c)(3) corporation whose mission is to help find missing children, reduce child sexual exploitation, and prevent child victimization.

Another important concern is the spread of Child Sexual Abuse Material (CSAM) online. CSAM refers to content that depicts sexually explicit activities involving a child.

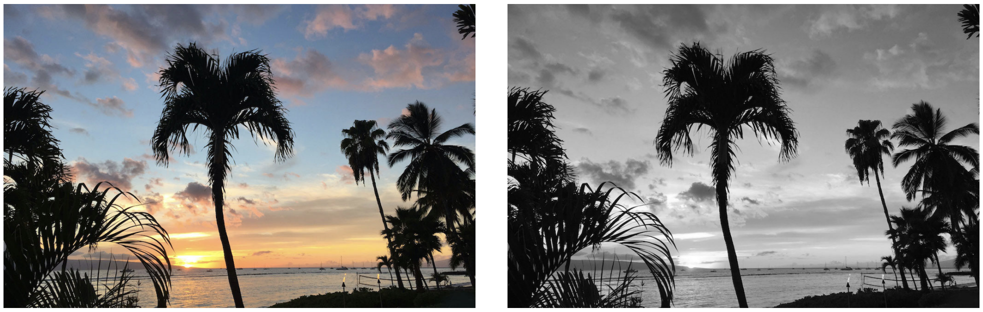
New technology in iOS and iPadOS[1] will allow Apple to detect known CSAM images stored in iCloud Photos. This will enable Apple to report these instances to the National Center for Missing and Exploited Children (NCMEC). NCMEC acts as a comprehensive reporting center for CSAM and works in collaboration with law enforcement agencies across the United States.

Apple's method of detecting known CSAM is designed with user privacy in mind. Instead of scanning images in the cloud, the system performs on-device matching using a database of known CSAM image hashes provided by NCMEC and other child safety organizations. Apple further transforms this database into an unreadable set of hashes that is securely stored on users' devices.


On Device
CSAM
Database of known CSAM Image Hashes
CSAM
List of Image Hashes
User Photos
Do they match?
Safety Vouchers
Uploaded to Apple
threshold not reached, no content visible
threshold reached, all content visible and unlocked
If threshold of matches is exceeded, only then can Apple interpret voucher content for CSAM matches
Reviewed by Apple, report sent to NCMEC

[1] Features available in the U.S.

The hashing technology, called NeuralHash, analyzes an image and converts it to a unique number specific to that image. The main purpose of the hash is to ensure that identical and visually similar images result in the same hash, while images that are different from one another result in different hashes. For example, an image that has been slightly cropped, resized or converted from color to black and white is treated identical to its original, and has the same hash.



NeuralHash: 100100111010100101...

NeuralHash: 100100111010100101...

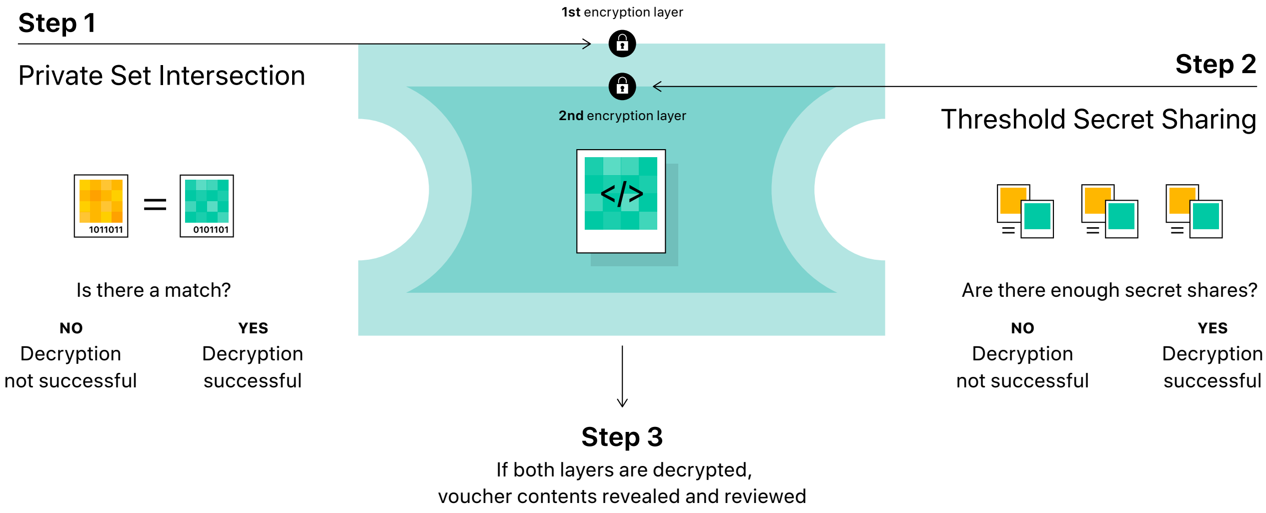
The image on the right is a black and white transformation of the image on the left. Because they are different versions of the same photo, they have the same NeuralHash.

Before an image is stored in iCloud Photos, an on-device matching process is performed for that image against the unreadable set of known CSAM hashes. This matching process is powered by a cryptographic technology called private set intersection, which determines if there is a match without revealing the result. Private set intersection (PSI) allows Apple to learn if an image hash matches the known CSAM image hashes, without learning anything about image hashes that do not match. PSI also prevents the user from learning whether there was a match.

The device then creates a cryptographic safety voucher that encodes the match result. It also encrypts the image's NeuralHash and a visual derivative. This voucher is uploaded to iCloud Photos along with the image. Using another technology called threshold secret sharing, the system ensures the contents of the safety vouchers cannot be interpreted by Apple unless the iCloud Photos account crosses a threshold of known CSAM content.

Threshold secret sharing is a cryptographic technique that enables a secret to be split into distinct shares so the secret can then only be reconstructed if the number of available shares exceeds a predefined threshold. For example, if a secret is split into one thousand shares, and the threshold is ten, then the secret can be reconstructed from any ten of the one thousand shares. However, if only nine shares are available, then nothing is revealed about the secret.

Only when the threshold is exceeded does the cryptographic technology allow Apple to interpret the contents of the safety vouchers associated with the matching CSAM images. Apple then manually reviews each report to confirm there is a match, disables the user's account, and sends a report to NCMEC. The threshold is set to provide an extremely high level of accuracy that accounts are not incorrectly flagged. This is further mitigated by a manual review process where Apple reviews each report to confirm there is a match. If so, Apple will disable the user's account and send a report to NCMEC. If a user feels their account has been mistakenly flagged they can file an appeal to have their account reinstated.



This innovative new technology allows Apple to provide valuable and actionable information to NCMEC and law enforcement regarding the proliferation of known CSAM. And it does so while providing significant privacy benefits over existing techniques since Apple only learns about users' photos if they have a collection of known CSAM in their iCloud Photos account. Even in these cases, Apple only learns about images that match known CSAM.

This design means that:

- This system is an effective way to identify known CSAM stored in iCloud Photos accounts while protecting user privacy.
- As part of the process, users also can't learn anything about the set of known CSAM images that is used for matching. This protects the contents of the database from malicious use.
- The system is very accurate, with an extremely low error rate of less than one in one trillion account per year.
- The system is significantly more privacy-preserving than cloud-based scanning, as it only reports users who have a collection of known CSAM stored in iCloud Photos.

Expanding guidance in Siri and Search

Apple is also expanding guidance in Siri and Search by providing additional resources to help children and parents stay safe online and get help with unsafe situations. For example, users who ask Siri how they can report CSAM or child exploitation will be pointed to resources for where and how to file a report.

Siri and Search are also being updated to intervene when users perform searches for queries related to CSAM. These interventions will explain to users that interest in this topic is harmful and problematic, and provide resources from partners to get help with this issue.

These updates to Siri and Search are coming later this year in an update to iOS 15, iPadOS 15, watchOS 8, and macOS Monterey.[1]

[1] Features available in the U.S.