Security Threat Model Review of Apple's Child Safety Features

August 2021

Communications safety in Messages

Feature goal

To enable parents to play a more informed role in helping their children navigate communication online through on-device interventions to children on their devices when sexually explicit images are sent or received. For children younger than 13, enable the parent to receive notification each time the child confirms these events.

Feature recap

For child accounts set up in Family Sharing, when the parent or guardian account opts in to the feature, a local, on-device machine learning classifier in the Messages app will warn the child if they are about to view or send sexually explicit images and will ask if they are sure they want to proceed. There are two warnings each time, each asking the child whether they wish to proceed; if the child is younger than 13 and the parent has opted in to receiving notifications for this feature, the second warning additionally lets the child know that proceeding to view the image will send a notification to their parents. The notification to parents does not contain the image in question. On devices associated with adult iCloud accounts, the only part of this feature that can be active is the ability to receive notifications from child accounts.

Design principles

The system is designed so that a user remains in control over all communication in the Messages app. This is achieved by making the feature opt-in at the parent/guardian level, requiring child accounts for which this feature is enabled to have set up Family Sharing, ensuring all child accounts have notice if the feature is enabled, and preventing notification to the parent until the child confirms each time that they wish to proceed with viewing or sending a sexually explicit image. Additionally, Apple gains no knowledge about the communications of any users with this feature enabled, and gains no knowledge of child actions or parental notifications.

Security and privacy requirements

We formalize the above design principles into the following security and privacy requirements.

- **Transparency:** the user must know when the system is enabled and whether the system will send a notification.

- **Consent:** the user must always have a choice about whether to take an action that can result in a notification being sent.

- **Control:** users who are parents or guardians of child accounts determine the status of this feature for those child accounts, older children have a path to disable the feature for their account, and the feature cannot be enabled for adult accounts.

- **Confidentiality:** the only information this feature can send from a child's device is a notification to the parents or guardians, with the child's confirmation. This feature will not send information to any other party.

Threat model considerations

This feature does not reveal information to Apple. Specifically, it does not disclose the communications of the users, the actions of the child, or the notifications to the parents. It does not compare images to any database, such as a database of CSAM material. It never generates any reports for Apple or law enforcement.

This feature cannot be enabled for an adult account, even with physical access to the device. It can only be activated by a parent/guardian account for a child account that's part of Family Sharing. The feature is not enabled for such child accounts by default.

If the feature were enabled surreptitiously or maliciously – for example, in the Intimate Partner Surveillance threat model, by coercing a user to join Family Sharing with an account that is configured as belonging to a child under the age of 13 – the user would receive a warning when trying to view or send a sexually explicit image. If they chose to proceed, they would be given a second warning letting them know that viewing the image will result in a notification being sent, and giving them another choice about

whether to proceed. If they declined to proceed, neither the fact that the warnings were presented, nor the user's decision to cancel, are sent to anyone.

The machine learning classifier used for this feature ships as part of the signed operating system. It is never downloaded or updated separately over the Internet or through any other mechanism. This claim is subject to code inspection by security researchers like all other iOS device-side security claims.

Because no data automatically or silently leaves the device, this feature significantly mitigates risk from false positives in the machine learning classifier, or from adversarial attacks against it. No false positives or adversarial ML attacks are possible against adult accounts, as the feature (and therefore the classifier) are disabled for such accounts.

For a child between the ages of 13–17, and whose account is opted in to the feature, sending them an adversarial image that the classifier misclassifies as sexually explicit, or sending them an image that incorrectly triggers a false positive classification means that, should the child decide to view the image, they will see something that's not sexually explicit.

For a child under the age of 13 whose account is opted in to the feature, and whose parents chose to receive notifications for the feature, sending the child an adversarial image or one that benignly triggers a false positive classification means that, should they decide to proceed through both warnings, they will see something that's not sexually explicit, and a notification will be sent to their parents. Because the photo that triggered the notification is preserved on the child's device, their parents can confirm that the image was not sexually explicit.

The same considerations apply to images being sent rather than received.

CSAM detection

Feature goal

This feature is designed to detect collections of illegal, known CSAM images stored on Apple servers in iCloud Photos libraries, while not learning any information about non-CSAM images.

Feature recap

For iCloud accounts which use iCloud Photos, this feature implements a privacy-preserving, hybrid on-device/server pipeline to detect collections of CSAM images being uploaded to iCloud Photos. The first phase runs code on the device to perform a blinded perceptual hash comparison of each photo being uploaded to iCloud Photos against an on-device encrypted database of known CSAM perceptual hashes. However, the result of each blinded match is not known to the device; it can only be determined by the second phase running on iCloud Photos servers, and only if that user's iCloud Photos account exceeds a threshold of positive matches.

The on-device encrypted CSAM database contains only entries that were independently submitted by two or more child safety organizations operating in separate sovereign jurisdictions, i.e. not under the control of the same government. Mathematically, the result of each match is unknown to the device. The device only encodes this unknown and encrypted result into what is called a safety voucher, alongside each image being uploaded to iCloud Photos. The iCloud Photos servers can decrypt the safety vouchers corresponding to positive matches if and only if that user's iCloud Photos account exceeds a certain number of matches, called the match threshold.

Before the threshold is exceeded, the cryptographic construction does not allow Apple servers to decrypt any match data, and does not permit Apple to count the number of matches for any given account. After the threshold is exceeded, Apple servers can only decrypt vouchers corresponding to positive matches, and the servers learn no information about any other images. The decrypted vouchers allow Apple servers to access a visual derivative – such as a low-resolution version – of each matching image.

These visual derivatives are then examined by human reviewers who confirm that they are CSAM material, in which case they disable the offending account and refer the account to a child safety organization – in the United States, the National Center for Missing and Exploited Children (NCMEC) – who in turn works with law enforcement on the matter.

Design principles

The system is designed so that a user need not trust Apple, any other single entity, or even any set of possibly-colluding entities from the same sovereign jurisdiction (that is, under the control of the same government) to be confident that the system is functioning as advertised. This is achieved through several interlocking mechanisms, including the intrinsic auditability of a single software image distributed worldwide for execution on-device, a requirement that any perceptual image hashes included in the on-device encrypted CSAM database are provided independently by two or more child safety organizations from separate sovereign jurisdictions, and lastly, a human review process to prevent any errant reports.

Security and privacy requirements

We formalize the above design principles into the following security and privacy requirements.

- **Source image correctness:** there must be extremely high confidence that only CSAM images – and no other images – were used to generate the encrypted perceptual CSAM hash database that is part of the Apple operating systems which support this feature.

- **Database update transparency:** it must not be possible to surreptitiously change the encrypted CSAM database that's used by the process.

- **Matching software correctness:** it must be possible to verify that the matching software compares photos only against the encrypted CSAM hash database, and against nothing else.

- **Matching software transparency:** it must not be possible to surreptitiously change the software performing the blinded matching against the encrypted CSAM database.

- **Database and software universality:** it must not be possible to target specific accounts with a different encrypted CSAM database, or with different software performing the blinded matching.

- **Data access restriction:** the matching process must only reveal CSAM, and must learn no information about any non-CSAM image.

- **False positive rejection:** there must be extremely high confidence that the process will not falsely flag accounts.

Threat model considerations

This feature runs exclusively as part of the cloud storage pipeline for images being uploaded to iCloud Photos and cannot act on any other image content on the device. Accordingly, on devices and accounts where iCloud Photos is disabled, absolutely no images are perceptually hashed. There is therefore no comparison against the CSAM perceptual hash database, and no safety vouchers are generated, stored, or sent anywhere.

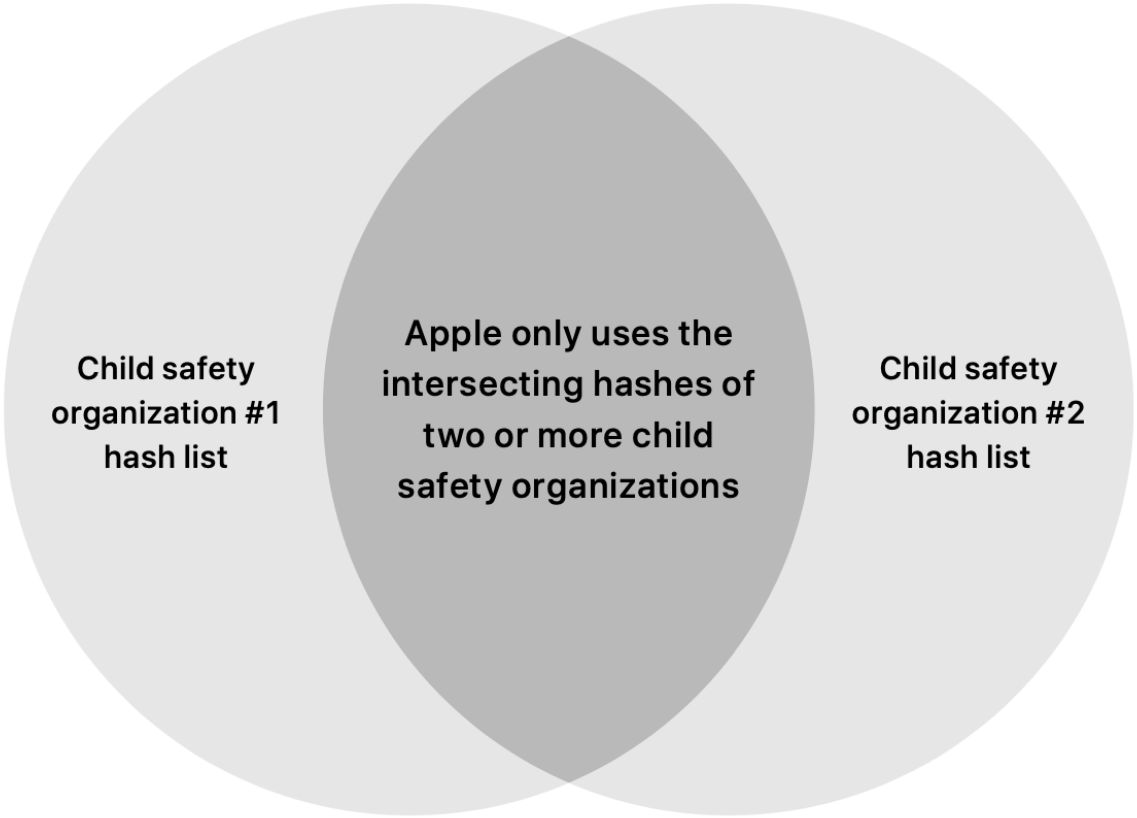
To discuss the potential for misuse of this system on devices where it is active – where the signed-in account uses iCloud Photos – we will examine the functioning of the system step by step.

Generation of the perceptual CSAM hash database

In the United States, NCMEC is the only non-governmental organization legally allowed to possess CSAM material. Since Apple therefore does not have this material, Apple cannot generate the database of perceptual hashes itself, and relies on it being generated by the child safety organization.

As Apple cannot inspect or verify this process, the threat model explicitly assumes the possibility of non-CSAM images being included in the perceptual CSAM hash database provided to Apple: either inadvertently, such as through an error at a designated child safety organization, or maliciously, such as through coercion. There are two strong protections against this risk.

The first protection against mis-inclusion is technical: Apple generates the on-device perceptual CSAM hash database through an intersection of hashes provided by at least two child safety organizations operating in separate sovereign jurisdictions – that is, not under the control of the same government. Any perceptual hashes appearing in only one participating child safety organization's database, or only in databases from multiple agencies in a single sovereign jurisdiction, are discarded by this process, and not included in the encrypted CSAM database that Apple includes in the operating system. This mechanism meets our source image correctness requirement.



The second protection is human review: there is no automated reporting in Apple's system. All positive matches must be visually confirmed by Apple as containing CSAM before Apple will disable the account and file a report with the child safety organization.

Distribution of the perceptual CSAM hash database

The perceptual CSAM hash database is included, in an encrypted form, as part of the signed operating system. It is never downloaded or updated separately over the Internet or through any other mechanism. This claim is subject to code inspection by security researchers like all other iOS device-side security claims.

Since no remote updates of the database are possible, and since Apple distributes the same signed operating system image to all users worldwide, it is not possible – inadvertently or through coercion – for Apple to provide targeted users with a different CSAM database. This meets our database update transparency and database universality requirements.

Apple will publish a Knowledge Base article containing a root hash of the encrypted CSAM hash database included with each version of every Apple operating system that supports the feature. Additionally, users will be able to inspect the root hash of the encrypted database present on their device, and compare it to the expected root hash in the Knowledge Base article. That the calculation of the root hash shown to the user in Settings is accurate is subject to code inspection by security researchers like all other iOS device-side security claims.

This approach enables third-party technical audits: an auditor can confirm that for any given root hash of the encrypted CSAM database in the Knowledge Base article or on a device, the database was generated only from an intersection of hashes from participating child safety organizations, with no additions, removals, or changes. Facilitating the audit does not require the child safety organization to provide any sensitive information like raw hashes or the source images used to generate the hashes – they must provide only a non-sensitive attestation of the full database that they sent to Apple. Then, in a secure on-campus environment, Apple can provide technical proof to the auditor that the intersection and blinding were performed correctly. A participating child safety organization can decide to perform the audit as well.

Matching against the perceptual CSAM hash database

Apple's perceptual hash algorithm, NeuralHash, is not an ML classifier. It has not been trained on CSAM images, e.g. to deduce whether a given unknown image may also contain CSAM. It does not contain extracted features from CSAM images (e.g. faces appearing in such images) or any ability to find such features elsewhere. Indeed, NeuralHash knows nothing at all about CSAM images. It is an algorithm designed to answer whether one image is really the same image as another, even if some image-altering transformations have been applied (like transcoding, resizing, and cropping).

We empirically assessed NeuralHash performance by matching 100 million non-CSAM photographs against the perceptual hash database created from NCMEC's CSAM collection, obtaining a total of 3 false positives, as verified by human inspection. Separately, we assessed NeuralHash on an adult pornography dataset of about 500,000 where we observed zero false positives against the perceptual hash database. The threat model explicitly takes into account the possibility of NeuralHash image-level false positives by imposing a required threshold of simultaneous matches for a given account before Apple's iCloud Photos servers can decrypt any vouchers.

Apple always chooses the match threshold such that the possibility of any given account being flagged incorrectly is lower than one in one trillion, under a very conservative assumption of the NeuralHash false positive rate in the field. As the system is initially deployed, we do not assume the 3 in 100M image-level false positive rate we measured in our empirical assessment. Instead, our conservative assumption builds in a safety margin that is two orders of magnitude stronger. Specifically, we assume a worst-case NeuralHash image-level error rate of one in one million, and pick a threshold that safely produces less than a one-in-one-trillion error rate for a given account under that assumption. Building in an additional safety margin by assuming that every iCloud Photo library is larger than the actual largest one, we expect to choose an initial match threshold of 30 images. Since this initial threshold contains a drastic safety margin reflecting a worst-case assumption about real-world performance, we may change the threshold after continued empirical evaluation of NeuralHash false positive rates – but the match threshold will never be lower than what is required to produce a one-in-one-trillion false positive rate for any given account.

The NeuralHash algorithm, the match threshold, and the cloud storage code that invokes NeuralHash to perform blinded matching on images being uploaded to iCloud Photos are all included as part of the code of the signed operating system. These components are never downloaded or updated separately over the Internet or through any other mechanism. This claim is subject to code inspection by security researchers like all other iOS device-side security claims. And since the code for these components is part of the operating system, security researchers can verify that it behaves as described, for example to confirm that the blinded matching code only matches against the on-device encrypted CSAM database and nothing else (meeting our matching software correctness requirement). And since Apple distributes the same signed operating system image to all users worldwide, it is not possible – inadvertently or through coercion – for Apple to provide targeted users with a different NeuralHash algorithm, match threshold, or blinded matching software. Taken together, these design and implementation properties meet our matching software transparency and software universality requirements.

Apple will publish, on the same Knowledge Base page containing the root hashes of perceptual CSAM hash databases shipped in Apple operating systems, the match threshold for each of those versions. Additionally, the user will be able to inspect the match threshold on their device.

For accounts using iCloud Photos, though the user's device performs the first phase of the perceptual hash comparisons against the encrypted CSAM database for photos being uploaded to iCloud Photos, the device is cryptographically prevented from knowing whether a match is successful or counting the number of positive matches. The device can therefore not check whether the threshold is exceeded, which means it will not – and cannot – report violating users to any entity, including Apple. The sole output of the blinded matching process are the safety vouchers for each image being uploaded to iCloud Photos, indicating in an encrypted way whether each specific voucher matches an entry in the perceptual CSAM hash database. The vouchers corresponding to positive matches can be decrypted by the server only when the match threshold is exceeded.

Match voucher decryption by iCloud Photos servers

Apple’s CSAM detection is a hybrid on-device/server pipeline. While the first phase of the NeuralHash matching process runs on device, its output – a set of safety vouchers – can only be interpreted by the second phase running on Apple's iCloud Photos servers, and only if a given account exceeds the threshold of matches. The local device does not know which images, if any, positively matched the encrypted CSAM database. iCloud Photos servers periodically perform the mathematical protocol to discover whether a given account exceeded the match threshold. With the initial match threshold chosen as described above, iCloud Photos servers learn nothing about any of the user's photos unless that user's iCloud Photos account exceeded the match threshold. This meets our data access restriction requirement.

To make sure Apple's servers do not have a count of matching images for users below the match threshold, the on-device matching process will, with a certain probability, replace a real safety voucher that's being generated with a synthetic voucher that only contains noise. This probability is calibrated to ensure the total number of synthetic vouchers is proportional to the match threshold. Crucially, these synthetic vouchers are a property of each account, not of the system as a whole. For accounts below the match threshold, only the user's device knows which vouchers are synthetic; Apple's servers do not and cannot determine this number, and therefore cannot count the number of true positive matches.

The code running on the device will never let Apple servers know the number of synthetic vouchers directly; this claim is subject to code inspection by security researchers like all other iOS device-side security claims. Only once an account exceeds the match threshold of true matches against the perceptual CSAM hash database can Apple servers decrypt the contents of the corresponding safety vouchers and obtain the exact number of true matches (always in excess of the match threshold) – and the visual derivatives that correspond to those vouchers. In other words, even though the creation of synthetic vouchers is a statistical protection mechanism, it is not a traditional noise-based approach: under this protocol, it is impossible for servers to distinguish synthetic vouchers from real ones unless the number of true positive (non-synthetic) matches cryptographically exceeds the match threshold.

Once the match threshold is exceeded, Apple servers can decrypt only the voucher contents that correspond to known CSAM images. The servers learn no information about any voucher that is not a positive match to the CSAM database. For vouchers that are a positive match, the servers do not receive a decryption key for their images, nor can they ask the device for a copy of the images. Instead, they can only access the contents of the positively-matching safety vouchers, which contain a visual derivative of the image, such as a low-resolution version. The claim that the safety vouchers generated on the device contain no other information is subject to code inspection by security researchers like all other iOS device-side security claims.

Human review and reporting

Once Apple's iCloud Photos servers decrypt a set of positive match vouchers for an account that exceeded the match threshold, the visual derivatives of the positively matching images are referred for review by Apple. First, as an additional safeguard, the visual derivatives themselves are matched to the known CSAM database by a second, independent perceptual hash. This independent hash is chosen to reject the unlikely possibility that the match threshold was exceeded due to non-CSAM images that were adversarially perturbed to cause false NeuralHash matches against the on-device encrypted CSAM database. If the CSAM finding is confirmed by this independent hash, the visual derivatives are provided to Apple human reviewers for final confirmation.

The reviewers are instructed to confirm that the visual derivatives are CSAM. In that case, the reviewers disable the offending account and report the user to the child safety organization that works with law enforcement to handle the case further.

The threat model relies on the technical properties of the system to guard against the unlikely possibility of malicious or coerced reviewers, and in turn relies on the reviewers to guard against the possibility of technical or human errors earlier in the system. A reviewer can only file a report for those accounts which have exceeded the match threshold, as analyzed by a technical system that the reviewers have no control over, as we described. And even if non-CSAM hashes were ever present in the on-device database (due to the hash set intersection, this would require mis-inclusion in the databases of at least two child safety organizations operating in separate sovereign jurisdictions), and this caused incorrect matches, or even in the mathematically unlikely case of a one-in-

one-trillion false identification of an account, a human reviewer would see that the visual derivatives for the purportedly-matching images are not CSAM, and would neither disable the account nor file a report with a child safety organization. Instead, they would refer the matter to engineering for analysis. The same would be true in the case of an adversarial attack against the NeuralHash perceptual algorithm, which could cause non-CSAM images to exceed the match threshold for a given account.

Since Apple does not possess the CSAM images whose perceptual hashes comprise the on-device database, it is important to understand that the reviewers are not merely reviewing whether a given flagged image corresponds to an entry in Apple's encrypted CSAM image database – that is, an entry in the intersection of hashes from at least two child safety organizations operating in separate sovereign jurisdictions. Instead, the reviewers are confirming one thing only: that for an account that exceeded the match threshold, the positively-matching images have visual derivatives that are CSAM. This means that if non-CSAM images were ever inserted into the on-device perceptual CSAM hash database – inadvertently, or through coercion – there would be no effect unless Apple's human reviewers were also informed what specific non-CSAM images they should flag (for accounts that exceed the match threshold), and were then coerced to do so. Surreptitious modifications to the CSAM hash database are therefore insufficient to cause innocent people to be reported with this system. Apple will refuse all requests to add non-CSAM images to the perceptual CSAM hash database; third party auditors can confirm this through the process outlined before. Apple will also refuse all requests to instruct human reviewers to file reports for anything other than CSAM materials for accounts that exceed the match threshold.