Council of the European Union
General Secretariat

Interinstitutional files:
2022/0155 (COD)

Brussels, 14 May 2024

WK 6847/2024 INIT

LIMITE

JAI	**FREMP**
ENFOPOL	**TELECOM**
CRIMORG	**COMPET**
IXIM	**MI**
DATAPROTECT	**CONSOM**
CYBER	**DIGIT**
COPEN	**CODEC**

This is a paper intended for a specific community of recipients. Handling and further distribution are under the sole responsibility of community members.

WORKING DOCUMENT

From:	Presidency
To:	Law Enforcement Working Party (Police)
Subject:	Proposal for a Regulation of the European Parliament and of the Council laying down rules to prevent and combat child sexual abuse - New approach proposed by the Presidency

Delegations are provided in the ANNEX with a Presidency note illustrating the new approach presented at the meeting of the Law Enforcement Working Party (Police) on 8 May 2024. Prior to a discussion of the new approach at an upcoming meeting, delegations are invited to submit comments by 22 May 2024 to CGI.LEWP.Be2024@police.belgium.eu and csa@consilium.europa.eu.

ANNEX

BE PCY new approach on targeting detection orders

At the LEWP meeting of 15 April 2024, the BE PCY concluded that its initial approach to the proposed CSA Regulation was insufficiently supported by the Member States. The BE PCY therefore presented a new approach at the LEWP meeting of 8 May 2024 that focuses on finding the right balance between proportionality and effectiveness and will hopefully sufficiently alleviate concerns of Member States regarding the protection of privacy as well as the protection of children's rights.

This note sets out the headlines of this new approach.

1. New scope: only visual content (images and videos)

Known CSAM, new CSAM and grooming would remain in the scope of detection orders.

However, detection would only be possible on visual content[1] and URLs. There would be no detection on audio and text.

Known CSAM, new CSAM and grooming would all remain in the scope of detection orders. Even without detection on audio and text, a fairly large part of grooming could still be detected. This is because during the grooming process there will almost always be an exchange of images or videos. Only in exceptional cases there is no exchange of images or videos. Grooming will therefore be identified through the detection of known and new CSAM.

Article 2 (y) is deleted (definition of real time audio communication and call).

Article 7 (3), second paragraph (b), (7) and (9) (last half-sentence of the last paragraph) are deleted (provisions related to detection orders for solicitation of children).

Article 44 (1) (c) is deleted (database for indicators to detect solicitation of children).

Recital 21 (second paragraph) and 28 (last sentence) are deleted (safeguards regarding the detection for solicitation of children).

ANNEX 1 is adapted (template for detection orders).

[1] "Images and visual components of videos" specifically refers to the graphical content of digital media. This includes photographs (images) and the visual sequences that make up videos, i.e., everything that can be visually represented and perceived in video content. This definition excludes any non-visual data in videos, such as audio tracks and embedded text data (such as subtitles or metadata). In this scenario, the detection technology would be applied only to the visual aspects of digital content to identify potential child sexual abuse material, without analysing or monitoring the audio components or any text data that might also be present in the video files.

New Recital 23a:

"To further avoid undue interference with fundamental rights and ensure proportionality, detection orders should cover only images and the visual components of videos and URLs, while the detection of audio communication and text should be excluded. Despite that limitation of detection to images and the visual components of videos, the solicitation of children could still be identified to some extent through the detection of visual material exchanged."

The use of detection technology will be different for known CSAM and new CSAM:

- Known CSAM: detection could be based on cryptographic and perceptual hashing, given the existing technologies.[2]

- New CSAM: detection could be based on artificial intelligence, such as machine learning tools, advanced algorithms, etc, given the existing technologies.[3]

To reduce false positives in the detection of new CSAM, additional safeguards will be included:

- Delayed reporting reducing false positives: 2 hits for new CSAM.

- New CSAM detected will be pseudonymized[4] prior to human verification. The pseudonymization will be applied to the metadata related to personal information, but even face blurring techniques could be considered. When there is a match of possible new CSAM, the provider would pseudonymize the metadata included in the report to the EU Centre, so that the EU Centre would not be able to know to whom the data belongs until it verifies that the content is not manifestly unfounded. In this case, the EU Centre would tell the provider and the provider would share with the EU Centre the metadata in the clear, so that the EU Centre can forward it to law enforcement.

[2] Cryptographic hashing and perceptual hashing are used for the detection of known CSAM. They create a unique fingerprint (hash) for each image they scan and compare this fingerprint against the hashes in a database (indicators) of known CSAM. However, cryptographic hashing is only able to detect two exact images. The added value of perceptual hashing, such as Photo DNA, Facebook's PDQ hash function and Apple's NeuralHash function, is that the generated hash is robust against common image transformations such as resizing, compression, blurring, noise, etc. while also being sensitive to perceptual similarity. .

[3] Machine learning algorithms are trained on large datasets of CSAM to learn patterns and features associated with CSAM. This however requires the analysis of the visual content of images and videos rather than hashing.

[4] With 'pseudonymisation' we refer to the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person (article 4, (5) GDPR).

Article 7 – possible new paragraph 6a (instead of paragraph 10, as this text relates only to the detection of new CSAM now):

"Providers of hosting services and providers of interpersonal communications services shall carry out the detection orders concerning the dissemination of new child sexual abuse material in a way that the material is reported in accordance with Articles 12 and 13 under the conditions outlined in sub-paragraphs 2 to 5.

The detection of potential new child sexual abuse material shall result in a hit to be flagged in the affected service, without the provider getting knowledge of, or control over, that information.[5] Providers shall preserve the information about the existence of the hit for at least twelve months or the duration of the respective detection order, whatever is longer.

A child using the service shall be automatically and immediately informed that potential new child sexual abuse material was detected, without the provider getting knowledge of, or control over, that information. The child shall be enabled to notify the provider thereof through tools that are easily accessible and age appropriate.

Once potential new child sexual abuse material has been flagged in a service twice or once a child has notified the provider about the detection of potential new child sexual abuse material within a service, the provider shall report that material to the EU Centre in such a manner that the personal data cannot be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separate and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.

Where the EU Centre considers, after human verification, that a report on potential new child sexual abuse material submitted by a provider is not manifestly unfounded, it shall require the provider to re-submit the report without the limitations outlined in sub-paragraph 4."

Amended Recital 23a:

"To further avoid undue interference with fundamental rights and ensure proportionality, the reporting by providers of hosting services and providers of publicly available interpersonal communications services on potential online new child sexual abuse material should be limited to that material either notified to them by a child or detected repeatedly on their services. The first detection of potential new child sexual abuse material on their services should be stored by the providers as a hit and preserved for at least twelve months or, if the detection order has been issued for a duration of more than twelve months, for the duration of a detection order. The stored hit should be deleted thereafter. As an additional safeguard, the reporting to the EU Centre of potential new child sexual abuse material detected in the service of a provider, should be done in a pseudonymized

[5] See Appendix 1.1. and Appendix 1.4. This is technically possible via technologies such as "zero-knowledge proofs" protocols (appendix 1.4.) and some alternatives "zero knowledge" processes were implemented with the coronavirus applications (Appendix 1.1).

way, so that the personal data cannot be attributed to a specific data subject. Only after human verification by the EU Centre, the providers should share the report with the EU Centre including the personal data attributable to a data subject."

2. No detection on E2EE data

The approach to end-to-end encryption stays the same: E2EE data remains outside the scope of detection orders and therefore protected from detection. However, detection of CSAM, including in services using E2EE, could be enabled via 'upload moderation' implying 'user consent'.

3. Detection subject to 'user consent'

The technology used to detect CSAM will be deployed via 'upload moderation'. Upload moderation allows for the detection to take place on the initiating end of the communication, in the case of services using E2EE before it has been encrypted. Upload moderation does not take place at the receiving end of the communication. The service provider will scan images and videos for CSAM (and URLs) before the user uploads them and sends them to the receiver, on the condition that the user gives its consent for such detection[6].

The technical process of how a service provider like WhatsApp® would go about this could be described as follows: for a user to send an image or video via WhatsApp to another user, WhatsApp needs access first to the user device's storage or camera and later to the user's device network to send the image or video to the other user. To do so, WhatsApp first needs permission from the user to access the storage or camera. Once it has permission, WhatsApp can use the operating system's (OS) application programming interface (API) to access the user's device storage and camera. The type of API depends on the OS (iOS, Android, …). Before WhatsApp uses the OS's API to establish a connection to the internet and send the images or videos and before it compresses and encrypts these images or videos, WhatsApp can scan these images or videos for the presence of CSAM.[7]

We propose to build in such 'user consent' by informing the user in the terms and conditions of the service provider of the fact that such detection could be deployed if the provider receives a detection order. The user would then have the choice:

1. To consent to the detection of visual content and URLs, in which case the user would have access to the entirety of the functionalities of that service, and notably to share visual content and URLs, or
2. To not consent to the detection of visual content and URLs, in which case the user would have access to all the functionalities of the service, except sharing visual content and URLs.

[6] See Appendix 1.3.
[7] See Appendix 1.2.

Possible amended Article 1(5):

"Without prejudice to Article 10a, this Regulation shall not prohibit or make impossible end-to-end encryption, implemented by the relevant information society services or by the users. This Regulation shall not create any obligation to decrypt or create access to end-to-end encrypted data, or that would prevent providers from offering end-to-end encrypted services."

Possible new paragraph 4(aa) in Article 10:

(aa) limit the functionalities of the service to prevent the transmission of visual content and URLs absent the user consent pursuant to paragraph 5(aa).

Article 10(4)(f) should include a reference to the new paragraph (i.e. "…regularly review the functioning of the measures referred to in points (a), (aa), (b), (c) and (d) of this paragraph…"

Possible amendments of Article 10(5):

5. The provider shall ~~inform users~~ **request the consent of users after informing them in the terms and conditions of use** in a clear, prominent and comprehensible way of the following:

   (a) the fact that, **upon receiving a detection order**, ~~it~~ **the provider** operates ~~**automated**~~ technologies ~~**(automated profiling)**~~ to detect online child sexual abuse~~,~~ to execute the detection order, the ways in which it operates those technologies, **meaningful information about the logic involved**, and the impact on the confidentiality of users' communications;

   (aa) the fact that, upon receiving a detection order in interpersonal communications services, it is required to limit the functionalities of the service to prevent the transmission of visual content and URLs absent the user consent;

   (b) the fact that **the provider** ~~it~~ is required to report potential online child sexual abuse to the EU Centre in accordance with Article 12;

   (c) the users' right of judicial redress referred to in Article 9(1) and their rights to submit complaints to the provider through the mechanism referred to in paragraph 4, point (d) and to the Coordinating Authority in accordance with Article 34.

   ~~**(d) the users' rights as data subjects under Regulation (EU) 2016/679.**~~

   The provider shall not provide information to users that may reduce the effectiveness of the measures to execute the detection order.

New Article 10a

"In order to implement this Regulation, providers of interpersonal communications services shall install and operate technologies to detect, prior to transmission, the dissemination of known child sexual abuse material or of new child sexual abuse material . "

New Recital 26a:

"While end-to-end encryption is a necessary means of protecting fundamental rights and the digital security of governments, industry and society, the European Union needs to ensure the effective prevention of and fight against serious crime such as child sexual abuse. Providers should therefore not be obliged to prohibit or make impossible end-to-end encryption. Nonetheless, it is crucial that services employing end-to-end encryption do not inadvertently become secure zones where child sexual abuse material can be shared or disseminated without possible consequences. Therefore, child sexual abuse material should remain detectable in all interpersonal communications services through the application of vetted technologies, when uploaded, under the condition that the users give their explicit consent under the provider's terms and conditions for a specific functionality being applied to such detection in the respective service. Users not giving their consent should still be able to use that part of the service that does not involve the sending of visual content and URLs. This ensures that the detection mechanism can access the data in its unencrypted form for effective analysis and action, without compromising the protection provided by end-to-end encryption once the data is transmitted."

Amended Article 50(1):

The EU Centre shall make available technologies that providers of hosting services and providers of interpersonal communications services may acquire, install and operate, free of charge, where relevant subject to reasonable licensing conditions, to execute detection orders in accordance with Article 10(1) and 10a.

Appendix 1.1 How the coronavirus application works and interacts with iOS

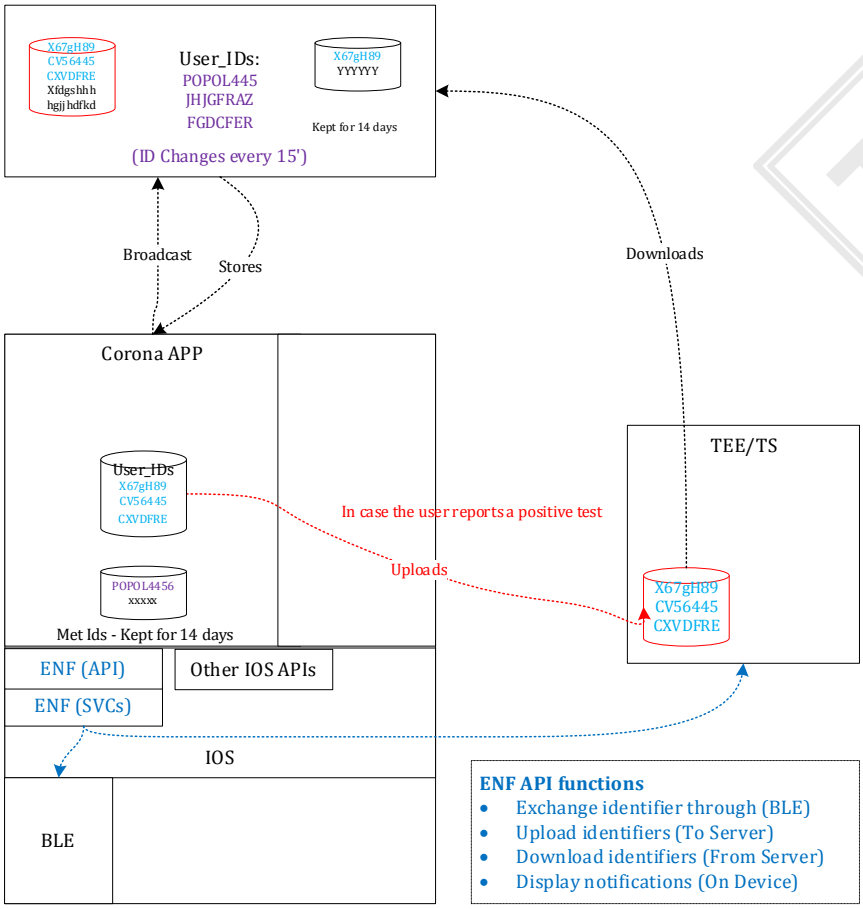
The coronavirus application is a mobile app that helps users track their exposure to COVID-19 and get notified if they have been in contact with someone who tested positive. The app also provides users with health advice and guidance based on their risk level and symptoms. The app is compatible with iOS devices and uses the **Exposure Notification framework** developed by Apple and Google to exchange anonymous identifiers with other nearby devices that have the app installed.

The app works by generating a **random and unique identifier for each user**, which changes every 15 minutes. The identifier is not linked to any personal or location information and cannot be used to identify the user. The **app broadcasts the identifier via Bluetooth Low Energy (BLE) to other nearby devices that have the app installed**. The app also scans for the identifiers of other users and stores them locally on the device for 14 days.

When a user reports a positive test result for COVID-19 in the app, the app uploads their identifiers from the past 14 days to a secure server. The server then distributes the identifiers to other devices that have the app installed. The app then compares the identifiers received from the server with the ones stored locally on the device. If there is a match, it means that the user has been in close contact with someone who tested positive for COVID-19. The app then notifies the user and provides them with health advice and guidance based on their exposure level and symptoms.

The app interacts with iOS through the **Exposure Notification framework**, which is a set of APIs and services that enable the app to use BLE for exchanging identifiers, to access the secure server for uploading and downloading identifiers, and to display notifications and alerts to the user. The framework also ensures that the app respects the user's privacy and consent, and that the app does not drain the battery or affect the performance of the device.

The app requires the user to **enable the Exposure Notification feature** in the Settings app and to grant the app permission to use BLE and to send notifications. The user can also disable the feature or revoke the permission at any time. The app does not collect or access any personal or location information from the user or the device. The **app only uses the identifiers generated by the framework, which are encrypted and anonymized**. The app also does not share the identifiers with any third parties, except for the secure server that is managed by the health authorities.

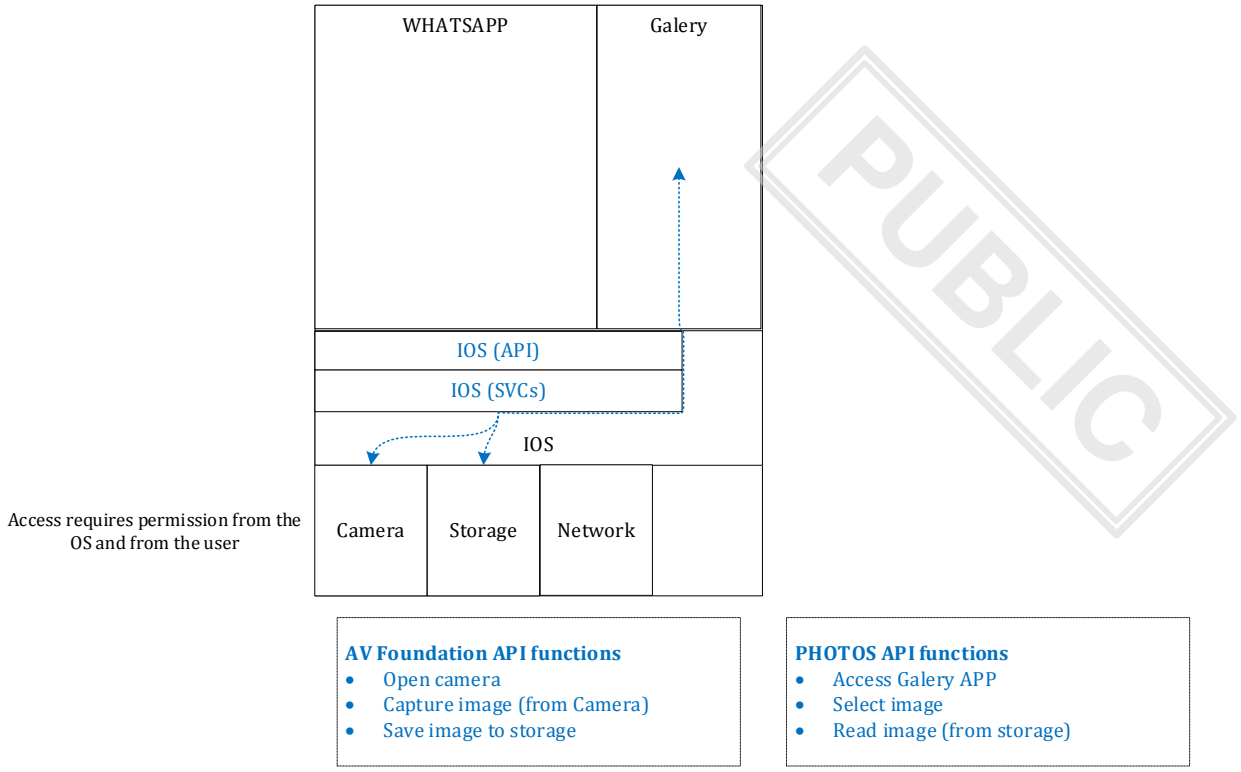

X67gH89
CV56445
CXVDFRE
Xfdgshh h
hgjj hdfkd
User_IDs:
POPOL445
JHJGFRAZ
FGDCFER
X67gH89
YYYYYY
Kept for 14 days
(ID Changes every 15')
Broadcast
Stores
Downloads
Corona APP
User_IDs
X67gH89
CV56445
CXVDFRE
In case the user reports a positive test
Uploads
POPOL4456
xxxxx
Met Ids - Kept for 14 days
TEE/TS
X67gH89
CV56445
CXVDFRE
ENF (API)
Other IOS APIs
ENF (SVCs)
IOS
BLE
ENF API functions
Exchange identifier through (BLE)
Upload identifiers (To Server)
Download identifiers (From Server)
Display notifications (On Device)

Appendix 1.2 Example: How to send a picture through WhatsApp

WhatsApp is a popular messaging app that uses the Extensible Messaging and Presence Protocol (XMPP) to send and receive text, voice, video, and image messages. One of the features of WhatsApp is that it can compress and encrypt the image files before sending them, reducing the data usage and ensuring the privacy of the users. In this document, we will explain the technical mechanisms of how WhatsApp communicates with the operating system (OS) and the device to get the picture and send it to the recipient.

When a user wants to send a picture through WhatsApp, they can either choose an existing picture from their gallery or take a new picture using the camera. In both cases, WhatsApp needs to access the device's storage and camera, which are controlled by the OS. To do this, WhatsApp needs to have the appropriate permissions from the user and the OS. Permissions are the rules that determine what an app can or cannot do on a device. For example, WhatsApp cannot access the camera without the user's consent and the OS's approval. When the user grants the permissions, WhatsApp can use the OS's application programming interface (API) to communicate with the device's hardware and software. An API is a set of commands and protocols that allow different programs to interact with each other. For example, WhatsApp can use the OS's API to request the device to open the camera app, capture the image, and save it to the storage. Similarly, WhatsApp can use the OS's API to access the gallery app, select the image, and read it from the storage.

Depending on the OS, WhatsApp may use different APIs to access the device's storage and camera. For example, on Android, WhatsApp may use the Storage Access Framework (SAF) to access the device's internal or external storage, and the Camera2 API to access the device's camera. On iOS, WhatsApp may use the Photos Framework to access the device's photo library, and the AV Foundation Framework to access the device's camera.



Once WhatsApp has the picture, it needs to send it to the recipient. To do this, WhatsApp needs to connect to the internet, which is also controlled by the OS. WhatsApp needs to have the permission to use the device's network, which can be either cellular or Wi-Fi. WhatsApp can use the OS's API to request the device to establish a connection to the internet and send the data packets.

Before sending the picture, WhatsApp compresses and encrypts it. Compression is the process of reducing the size of the file by removing some of the redundant or unnecessary information. Encryption is the process of transforming the file into a secret code that can only be decoded by the intended recipient. Compression and encryption help WhatsApp to save bandwidth, speed up the transmission, and protect the user's privacy.

WhatsApp uses the JPEG compression algorithm to compress the image files, which reduces the quality of the image but also the file size. WhatsApp uses the AES-256 encryption algorithm to encrypt the image files, which uses a symmetric key to encrypt and decrypt the data. The key is generated by WhatsApp and shared with the recipient using the Diffie-Hellman key exchange protocol, which is a secure way of exchanging keys over a public network.

After compressing and encrypting the picture, WhatsApp sends it to the WhatsApp server, which is a computer that hosts the WhatsApp service. The WhatsApp server receives the picture and forwards it to the recipient's device, where it is decrypted and decompressed by the WhatsApp app. The recipient can then view the picture on their device.

Appendix 1.3 High-level message sending process

To detect images in the context of sending messages only, the scan of the images is to be triggered.

- As part of a message creation flow started from the application
- By the OS that needs to act as a relay (The scanning of the image is done by a trusted application, but it is triggered by the messaging application, then by the OS)
- The trusted APP is invoked by an online or offline (message-bus) communication – The online approach was used as an example here.

2 high-level workflows are shown here to show the segregation of duties and trusted required between parties. The workflows cover 2 use-cases.

Use-case 1: Import an existing picture from the gallery (from the messaging application)

As an example, the scanning could be implemented into the get image from storage IOS function which triggers a call either directly to the scanning application or indirectly through a message BUS. If the encrypted image or its hash have been obtained by the scanning application and stored in a safe place, the sending of the message can continue.

Messaging APP	OS	Scanning APP
Create message (Messaging APP)		
Add image from Galery		
API CALL	**Access gallery**	
API CALL	**Select image**	
API CALL	**Read image**	
	Get image from storage	
	API CALL	**Scan image (In trusted APP)**
		Store image on secure enclave
Send message	Return image	Hash and encrypt image
		Homomorphic function → Result
		IF CSAM detected trigger next steps of procedure

Remarks

- There are **no modifications** for the messaging APP
- The encryption of the image could also be the responsibility of the OS provider using the homomorphic encryption related public keys stored in the secure enclave (in the example provided it is the responsibility of the third party)
- Other mechanisms can be explored like the possibility to monitor for changes and triggers like changes on the gallery or on the usage of the camera - To stay in the message creation context the trigger needs to come from a message creation flow but we could imagine the OS sending a message to a message bus instead of invoking the secured application directly.

Use-case 2: Take a picture from the camera in the messaging application

In the second use-case, the scanning could be implemented into the get image from camera IOS function.

Messaging APP	OS	Scanning APP
Create message (Messaging APP)		
Add image from Camera		
API CALL	**Open camera**	
API CALL	**Capture Image**	
	Get image from camera	
	API CALL	**Scan image**
		Store image on secure enclave
	Save image to storage	Hash and encrypt image
Send message		Homomorphic function → Result
		IF CSAM detected trigger next steps of procedure

Appendix 1.4 Zero-knowledge proofs (ZPKs)

Zero-knowledge proofs (ZKPs) are a cutting-edge technology that is gaining attention in various fields, particularly in the area of blockchain. ZKPs allow for the verification of information without revealing the underlying data, providing a high level of security and privacy. In this article, I'll explore the basics of ZKPs, including how they work, why they are essential and their application in the blockchain.

Zero-knowledge proofs (ZKPs) enable the validation of a claim without disclosing any details about the statement in question. The concept of ZKPs was first introduced in a 1985 paper titled "The knowledge complexity of interactive proof systems" by Shafi Goldwasser, Silvio Micali and Charles Rackoff.

In a ZKP, two parties are involved: the prover and the verifier. The prover aims to establish a claim, and the verifier is accountable for verifying the claim. The prover can demonstrate to the verifier that a statement is accurate without revealing any supplementary information regarding the statement. This is done by providing proof, or a small amount of information that can be verified by the verifier to ensure that the statement is true.

One example of a ZKP is a proof of knowledge, where the prover demonstrates that they know a specific value without revealing the value itself. Another example is a proof of identity, where the prover can prove their identity without revealing any personal information.

The definition of a zero-knowledge protocol: "A zero-knowledge protocol is a method by which one party (the prover) can prove to another party (the verifier) that something is true, without revealing any information apart from the fact that this specific statement is true."

The most attractive feature of zero-knowledge proof lies in its seemingly contradictory unique nature that a prover can prove the correctness of an assertion to the verifier without leaking any extra information. It can force the malicious participants in cryptographic protocol to execute in accordance with predetermined steps to ensure the safety of the protocol. Thus, it has a broad application prospect. To speak vividly, a verifier who receives the zero-knowledge proof of a statement is supposed to be told by God that it is true. The main features of zero-knowledge proof system include completeness, soundness, and zero knowledge.