Department of Information Engineering and
Computer Science

Privacy & Intellectual Property Rights

# ChatControl 2.0

TODO: write subtitle

Riccardo Gennaro

riccardo.gennaro@studenti.unitn.it

August 14, 2024

# Contents

## Abstract

# 1 Introduction

# 2 Related work

## 2.1 Content scanning aglorithms

There are two main classes of algorithms used for content scanning: hash-based and ML-based algorithms[1].

### 2.1.1 Hash-based detection

A hash function is a deterministic function that takes an arbitrary-sized input and returns a fixed-size output. For example, hashing a picture using the hashing algorithm SHA-512 always produces a 512-bit string.

These functions are used since it is faster to compare two strings than two entire images. Furthermore, this prevents providers from storing illegal material, since only pre-computed hashes of the target content are needed[1].

Given this design, mapping every possible picture to a set with cardinality $2^{256}$ produces outputs that belong to multiple pre-images, i.e. collisions.

Furthermore, sharing media multiple times results in the message being lossy compressed various times, e.g. sharing a message using WhatsApp. This means that this shared media will now produce a different hash.

For these reasons, *perceptual hashing* is used in the context of hash-based detection[1]. Perceptual hashing algorithms are used to recognize targeted media even if it did undergo minor modifications since they produce the same hash even if the input has been modified, e.g. compressed, cropped[1].

In particular, Steinebach's analysis founds that with the lowest threshold PhotoDNA was able to recognise dissimilar images with an false-positive rate (collision) of only 0.3%.

An example of perceptual hashing algorithm is Microsoft's PhotoDNA. It has been proved that these algorithms work well and raises low false-positive alerts[16], albeit not in an adversarial environment[14].

### 2.1.2 ML-based detection

This second method uses classifiers to understand if the tested content can be considered targeted material. The main difference between ML detection and Hash detection is that the first is capable of recognizing also non-previously known material, while the latter bases its detection mechanism on hashes of previously known target material[1].

In particular, this type of technologies are able to detect target behaviour in both text-based communications and multimedia files.

Discussing child grooming in online conversations, Bours and Kulsrud (2019) in [3] evaluate different methods for detecting such behaviour using different types of classifiers, e.g. Neural Networks (NN) and Support Vector Machine (SVM).

Results for processing of single messages in conversations show that the highest precision (0.68) is found using Logistic Regression. Furthermore, processing all the messages of a single chatter foster higher precision in determining if a user is engaging in the target behaviour[3]. As stated in the state of art of [3], to obtain the most accurate results the entirety of the conversation nedd to be processed.

Regarding previously unknow CSAM, Hee-Eun Lee et al. report in their survey how studies between 2012 and 2018 reached varying accuracies trying to solve this classification problem [13]. In particular, Castrillon-Santana et al. (2018) was able to obtain a 7% error rate in classifying adult/no-adult classification [4], while, Vitorino et al. (2018) were able to assess new CSAM using Deep-Learning algorithms with a true positive rate of 87.2% and true negative rate of 85.0% [20].

## 2.2 Content scanning deployment

To scan an exchange of messages between a sender and a set of receivers the matching algorithm can be set to listen on server-side or on client-side. These two types of scan are called *server-side scanning (SSS)* and *client-side-scanning (CSS)*. Following, I will report on the difference between these two methods.

### Server side scanning

The description of this operation flow is based on [1]. In SSS, the messages are scanned by the server. More specifically, a matching algorithm is created by the service provider or it's shared by a third-party organization. Successively, hashes of targeted material or a trained classifier are shared by the targeted material provider to the service provider. At this point, the detection algorithm is deployed on the server side and it's used to scan the communication that transits through the server.

It is important to note that this mechanism will not work if the communication is *end-to-end encrypted (E2EE)* since the messages can be decrypted only by the sender and the receiver and not by the server. To solve this problem, client-side scanning is used.

### Client side scanning

As described in [1], CSS can solve the problem of detecting target material in E2EE communications since the analysis is performed on the client side. The operational flow is similar to the server-side counterpart with the exception that the database containing the hashes of the target material or the trained ML model is sent by the service provider to the clients that will use it to analyze the messages before being encrypted and sent (for the sender) and after being decrypted (for the receiver)[1]. After having detected some suspicious material, an alert and a copy of the media will be sent to the service provider for verification.

## 2.3 Digital privacy and general monitoring in the EU

In the European Economic Area (EEA), the use of these technologies to prevent, detect, and prosecute criminal offenses is strictly regulated. Following, I will summarize the most important pieces of EU law that sit at the intersection between EU privacy rights, general monitoring, and the use of these technologies.

### 2.3.1 EU Charter of Fundamental Rights

The European Union considers the right to one's privacy in (digital) communications as one of the fundamental ones. In particular, Art. 7 of the Charter of Fundamental Rights reads "Everyone has the right to respect for his or her private and family life, home and communications"[5].

Nonetheless, Art.52 §3 imposes the same limitations to these rights as the one specified in the *European Convention on Human Rights (ECHR)*[5]. In particular, Art. 8 §2, reads that "There shall be no interference by a public authority with the exercise of this right except such as is in accordance with the law and is necessary in a democratic society in the interests of national security, public safety or the economic well-being of the country, for the prevention of disorder or crime, for the protection of health or morals, or the protection of the rights and freedoms of others"[9], thus limiting the scope of these rights.

### 2.3.2 General Data Protection Regulation - 2016/679

While the *General Data Protection Regulation (GDPR)* with further modifications does not provide explicit provisions regarding general monitoring, it does specify, inter alia, the methods for processing both personal and sensitive data.

In particular, *personal data* is defined in Art. 4(1) of this regulation [11] as "any information relating to an identified or identifiable natural person". Meanwhile, with *sensitive data* are a set of data that belongs to those special categories that can be found in Art.9 §1[11].

Regarding the context of this study, Art. 5 §1 point c of the GDPR[11] specifies that the amount gathered personal data should be "adequate, relevant and limited to what is necessary [...]", thus establishing the so-called 'data minimization'.

Furthermore, as prescribed by Art.9 §2, processing sensitive data is limited, inter alia, to either user consent (point a) or a necessity for the public interest, but limited by the fundamental rights and by "essence of the right to data protection"[11] (point g).

### 2.3.3 ePrivacy Directive - 2002/58/EC

The ePrivacy Directive of 12 July 2002 lays down provisions and rights regarding the privacy of EU citizens concerning electronic communications. Relevant articles to this discussion are Art. 5 §1 and Art. 15 §1.

The first ensures the confidentiality of the communication since it "prohibit(s) listening, tapping, storage or other kinds of interception or surveillance of communications and the related traffic data by persons other than users" [6].

On the other hand, Art. 15 §1 lays down a set of exceptions that limits the scope of this right, i.e., inter alia, "prevention, investigation, detection and prosecution of criminal offenses" [6].

Nonetheless, these exceptions must still be in accordance with Art.6 §(1) and §(2) of the Treaty on the European Union, ergo, in turn, they must respect Art.7 and Art.8 of the Fundamental Rights of the EU [6] [5] [18].

### 2.3.4 Interim Regulation - 2021/1232

The Interim Regulation is a temporary regulation that entered into effect on the 2nd of August 2021 to combat online *child sexual abuse (CSA)*. The temporary measure of this regulation is because a legislative framework to fight this phenomenon while maintaning data privacy is still to be agreed upon.

In particular, it is stated in Art. 1 §1 of this regulation that the subject matter is laying down a set of derogations from several obligations specified in the ePrivacy directive previously discussed[15]. Most important is Art. 3 §1 which prescribes how Art. 5 §1 and 6 §1 will not apply anymore concerning the confidentiality of those communications where there is a well-founded suspicion of online CSA[15].

Finally, it is important to note that the Regulation does not impose a service provider to check communications for CSA but only allows them to perform it voluntarily as written in Art. 3 §1[15]. As per Art. 10, this regulation was set to apply until August 2024, but was extended to the apply until the 3rd of April 2026[15].

## 2.4 CSAM proposal - 2022/0155 (COD)

In this section, I'll introduce the CSAM proposal focusing on the obligations of the service providers and the authorities.

The CSAM Regulation proposal, also known as ChatControl Regulation, is a proposal for regulation set to find a permanent legislative framework to replace the temporary Interim Regulation previously discussed.

Furthermore, differently from the above-cited derogation, this proposal aims at forcing service providers to scan conversations for not only previously known CSAM but also newly produced material and text conversations that happen with the intent of grooming a minor, as stated in Chapter II concerning obligations of the service providers[8].

The three main actors in this Regulation are

- *Service Provider (SP)*: in this discussion referring to providers of hosting services and providers of interpersonal communication services as in Art. 3.
- *Coordinating Authority (CA)*: designated by the Member State, is responsible for applying and enforcing the Regulation in the concerned Member State as per Art. 25.
- *EU Centre (EUC)*: a new European Agency. Part of its obligations are to create and maintain a database of indicators of CSAM, e.g. hash fingerprints, to decide which technologies to adopt for detecting the target material, and to provide such technologies to the service providers (see Art. 44, and Art. 50).

Regarding obligations, service providers must develop a risk assessment to report on the risk related to CSAM spread and grooming on their platform or service as prescribed by Art. 3 [8]. Again in Art. 3, the proposal lays down a list of factors helpful in identifying such risks such as previous instances of the target behaviour, the extent to which the service is used by children, but also the presence or not of functionalities like enabling users to search other users and share multimedia files.

Successively, the report is sent to the Coordinating Authority of the state in which the service provider has its legal office as per Art. 5. If risk is identified by the CA as per Art. 7 §4, the authority will issue a detection order as defined by Art. 7 §1.

This order entails the entry in effect of the obligation for the interested service provider to apply mitigations for such risk. Moreover, the interested SP will be obliged to deploy a set of technologies with the intent of scanning conversations and file sharing between users to detect alleged CSAM sharing or grooming activities, and successively reporting them to the competent authorities (see Art.3, 4, 5) [8].

Furthermore, such technologies are required by Art.10 to be effective in detecting CSAM, not able to extract any other information by the communication that is not relevant to the purpose of the detection, the least intrusive and sufficiently reliable, i.e. not error-prone.

# 3 Methodology

# 4 Findings

## 4.1 Technical evaluation of the proposal suitability

The aim of the proposal is to detect and report communications in which CSAM is present. In particular, the types of CSAM to be recognised are known CSAM, unknown CSAM, and child sexual solicitation (grooming).

While the proposal does not focus on the implementation details, the impact assessment accompanying the proposition presents two technical annexes. Annex 8 explains technologies covered in section 2.1. Annex 9 evaluates different approaches for CSS and SSS, introduced in section 2.2.

### 4.1.1 Detection of known CSAM

To detect previosuly known CSAM in an E2EE communication, it is proposed in Annex 9 to use server side scanning implementing a system similar to Microsoft's PhotoDNA but hashing the shared content on client-side before the encryption that happens before the material is sent[7].

While it is stated in Annex 9, the proposal fails to recognise that perception hashing does not perform well in an adversarial context.

To elaborate on this, it is possible to generate images that will trigger false positives by injecting adversarial noise in a non-CSAM image to make hashing algorithm an hash that will collide with CSAM hash[1].

To avoid these types of attacks, the assessment aims at following a "security by obscurity" desing to avoid leacking of the algorithm. This type of design is not state of the art and is not secure by design by itself[19]. Furthermore, the algorithm will have to be run on client side, making maintaining the secrecy of the hasher implementation difficult[2].

### 4.1.2 Detection of unknown CSAM

As previosuly described, to detect unknown CSAM machine learning classifiers are needed. Discussing the context in which the communication is E2E encrypted, the impact assessment propose the deployment of the classifier model on client side and operating the classification previosly the encryption of the media. Furthermore, this approach is difficult to implement especially on smartphones and other devices with lower computational capabilities.

Morevoer, both the accompanying and the complementary impact assessment report that there exist a classifier for detecting new CSAM, Thorn's Safer, with a precision of 99.9% and a recall of 80%[7] [10]. Unfortunately, no proof is provided to support such claims and the citation linking to the benchmarking platform redirect to a benchmarking platform for perceptual hashing algorithms and not for classifiers of previously unknown CSAM[3].

Furthermore, the performance reported in these documents are optimistic at least if compared with what has been summarize in section 2.1.2, and also taking into consideration that the Thorn's model for CSAM detection is deployed on server side[17] having access to higher computing power.

[1] Such an attack can be performed against Apple's Neuralhash as shown in https://github.com/greentfrapp/apple-neuralhash-attack

[2] See https://github.com/AsuharietYgvar/AppleNeuralHash2ONNX

[3] The link to the benchmarking platform is (https://perception.thorn.engineering/en/latest/examples/benchmarking.html)

### 4.1.3 Sexual solicitation of children

Also in this case, ML classifiers are required to scan conversations and flag grooming behaviour. Again, in E2EE communications the model must be deployed client-side. Text-based classifiers are easier to implement client-side than models based on other media as also noted in Annex 9 of the complementary impact assessment[7]. Furthemore, server-side based solutions developed under project Arthemis are reported to have an accuracy of only 88%[7].

### 4.1.4 Avoiding detection during CSAM sharing

A simple way to avoid detection under this proposed solution is to encrypt the media before sharing on the controlled platform. This issue is also been noted in the EDPB-EDPS Joint Opinion[2] but nevera adressed in any other reviewed document.

To elaborate on this, two adversaries that aim at maintaining the file sharing anonymous could engage in a key exchange as in an insecure communication channel. Having computed a shared secret, such secret is used to compute a symmetric key used to encrypt the media.

The following bash code and figure show how two users can share media without the system being able to detect CSAM material.

```
# Generate EC parameters for the prime256v1 (secp256v1)
      curve and save them to ecparam.pem
openssl ecparam -name prime256v1 -out ecparam.pem
# Generate a private key for Party 1 using the prime256v1
      curve and save it to party1_private_key.pem
openssl ecparam -name prime256v1 -genkey -noout -out
      party1_private_key.pem
# Extract and save the public key from Party 1's private
      key to party1_public_key.pem
openssl ec -in party1_private_key.pem -pubout -out
      party1_public_key.pem
# Derive a shared secret using Party 1's private key and
      Party 2's public key, saving the result to
      shared_secret.bin
openssl pkeyutl -derive -inkey party1_private_key.pem  -
      peerkey party2_public_key.pem -out shared_secret.bin
# Encrypt body.bin using AES-256-CBC with the shared secret
       (no salt used) and save the output to body.ecb.bin
openssl enc -aes-256-cbc -nosalt -pass file:./shared_secret
      .bin -in body.bin -out body.ecb.bin
```

Listing 1: Client 1 exchange a secret over unsecure channel and image encryption

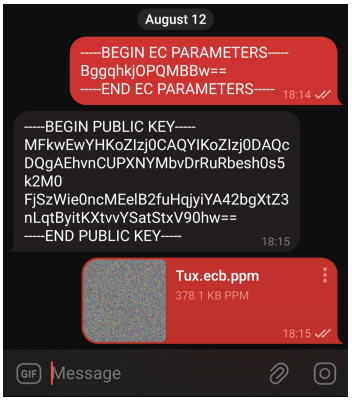


Figure 1: ECDH key exchange and AES encrypted image sharing

## 4.2 Relevant ECHR Cases

Following three cases discussed by the European Court of Human Rights about mass surveillance are presented. All three cases present discussions relating to violations in Articles 8 and 10 of the EU Charter of Fundamental Rights.

### 4.2.1 Big Brother et al. v. the United Kingdom

As described in [12], following the publication of the Snowden files that revealed the existence of an international surveillance system, the applicants complained an interference with their rights expressed by Article 8 and Article 10. The applicants claimed that their communications or communication data were tracked by the UK intelligence or obtained via service providers and/or via foreign intelligence.

The court ruled that there was indeed an interference with both Article 8 and Article 10 since secret bulk intercpetion was taking place. In particular, the Court defined bulk interception as a four stage process in which (a) there is the interception and initial

retention of communication data, (b) application of specific selectors to the retained data, e.g. queries, (c) examination of the selected data by analysts, (d) data retention.

Subsequently, the Court stated that, in a lawful bulk interception, the State should present clearly the grounds upon which the interception may take place.

# 5 Discussion

# 6 Conclusion

# References


[1] Harold Abelson et al. “Bugs in our pockets: The risks of client-side scanning”. In: *Journal of Cybersecurity* 10.1 (2024), tyad020. URL: https://www.researchgate.net/publication/355233857_Bugs_in_our_Pockets_The_Risks_of_Client-Side_Scanning.

[2] European Data Protection Board. *EDPB-EDPS Joint Opinion 04/2022 on the Proposal for a Regulation of the European Parliament and of the Council laying down rules to prevent and combat child sexual abuse*. 2022. URL: https://www.edpb.europa.eu/system/files/2022-07/edpb_edps_jointopinion_202204_csam_en_0.pdf (visited on 08/10/2024).

[3] Patrick Bours and Halvor Kulsrud. “Detection of Cyber Grooming in Online Conversation”. In: *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*. 2019, pp. 1–6. DOI: 10.1109/WIFS47025.2019.9035090.

[4] Modesto Castrillón-Santana et al. “Evaluation of local descriptors and CNNs for non-adult detection in visual content”. In: *Pattern Recognition Letters* 113 (2018). Integrating Biometrics and Forensics, pp. 10–18. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2017.03.016. URL: https://www.sciencedirect.com/science/article/pii/S0167865517300922.

[5] *Charter of Fundamental Rights of the European Union*. Official Journal of the European Union. Charter. European Union, Oct. 26, 2012. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:12012P/TXT (visited on 08/04/2024).

[6] *Directive on privacy and electronic communications*. Official Journal of the European Union. Directive. European Union. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02002L0058-20091219 (visited on 08/03/2024).

[7] European Commission. *IMPACT ASSESSMENT REPORT Accompanying the document Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL laying down rules to prevent and combat child sexual abuse*. European Commission. 2022. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52022SC0209 (visited on 08/10/2024).

[8] European Commission. *Proposal for a Regulation of the European Parliament and of the Council laying down rules to prevent and combat child sexual abuse*. European Commission. COM(2023) XXX final. 2023. URL: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12726-Fighting-child-sexual-abuse-detection-removal-and-reporting-of-illegal-content-online_en (visited on 08/06/2024).

[9] *European Convention on Human Rights*. Official Journal of the European Union. Convention. Council of Europe, Nov. 4, 1950. URL: https://www.echr.coe.int/Documents/Convention_ENG.pdf (visited on 08/05/2024).

[10] European Parliament. *Proposal for a regulation laying down the rules to prevent and combat child sexual abuse Complementary impact assessment*. European Parliament. 2023. URL: https://www.europarl.europa.eu/RegData/etudes/STUD/2023/740248/EPRS_STU(2023)740248_EN.pdf (visited on 08/11/2024).

[11] *General Data Protection Regulation*. Official Journal of the European Union. Regulation. European Union. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02016R0679-20160504 (visited on 08/01/2024).

[12] European Court of Human Rights. *CASE OF BIG BROTHER WATCH AND OTHERS v. THE UNITED KINGDOM*. 2021. URL: https://hudoc.echr.coe.int/eng#%7B%22itemid%22:%5B%22001-210077%22%5D%7D (visited on 08/11/2024).

[13] Hee-Eun Lee et al. "Detecting child sexual abuse material: A comprehensive survey". In: *Forensic Science International: Digital Investigation* 34 (2020), p. 301022. ISSN: 2666-2817. DOI: https://doi.org/10.1016/j.fsidi.2020.301022. URL: https://www.sciencedirect.com/science/article/pii/S2666281720301554.

[14] Jonathan Prokos et al. *Squint Hard Enough: Evaluating Perceptual Hashing with Machine Learning*. Cryptology ePrint Archive, Paper 2021/1531. https://eprint.iacr.org/2021/1531. 2021. URL: https://eprint.iacr.org/2021/1531.

[15] *Regulation on Temporary Derogation from Certain Provisions of the ePrivacy Directive for the Purpose of Combatting Child Sexual Abuse Online*. Official Journal of the European Union. Regualtion. European Union. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02021R1232-20240515 (visited on 08/03/2024).

[16] Martin Steinebach. "An analysis of photodna". In: *Proceedings of the 18th International Conference on Availability, Reliability and Security*. 2023, pp. 1–8. URL: https://dl.acm.org/doi/abs/10.1145/3600160.3605048.

[17] Thorn. *How Thorn's CSAM classifier uses artificial intelligence to build a safer internet*. 2023. URL: https://www.thorn.org/blog/how-thorns-csam-classifier-uses-artificial-intelligence-to-build-a-safer-internet/ (visited on 08/10/2024).

[18] *Treaty on European Union*. Official Journal of the European Union. Treaty. European Union, Oct. 26, 2012. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A12012M%2FTXT (visited on 08/05/2024).

[19] M. Aibin V. Fulber-Garcia. *Understanding Security by Obscurity*. 2024. URL: https://www.baeldung.com/cs/security-by-obscurity (visited on 08/10/2024).

[20] Paulo Vitorino et al. "Leveraging deep neural networks to fight child pornography in the age of social media". In: *Journal of Visual Communication and Image Representation* 50 (2018), pp. 303–313. ISSN: 1047-3203. DOI: https://doi.org/10.1016/j.jvcir.2017.12.005. URL: https://www.sciencedirect.com/science/article/pii/S1047320317302377.