Balancing Child Safety and Privacy:
An Exploration of the EU's Chatcontrol Proposal

Riccardo Gennaro
Department of Information Engineering and Computer Science
University of Trento
Via Sommarive, 9
Trento, TN 38123, Italy

This work explores the recent legislative proposal aimed at combating the proliferation of *Child Sexual Abuse Material (CSAM)* and online child sexual solicitation, also referred to as grooming. The proposal adopted by the European Commission suggests employing bulk interception of communication and the use of machine learning classifiers and hashing algorithms to detect and report such criminal behaviors. This solution raised significant concerns regarding privacy rights and the future of *End-to-End Encryption (E2EE)*.

This paper examines the technical feasibility, efficiency, and efficacy of the technological solutions also considering their deployment in an adversarial context. Furthermore, to estimate the impact on privacy rights, this discussion focuses on how this proposal relates to the current legislative framework by also reporting recent juridical cases to understand the interpretations given to privacy as a Fundamental Right.

Findings and subsequent discussion suggest that while child safety needs to be enhanced in digital contexts, the current way of implementing this safety could significantly harm privacy rights.

Ultimately, this work calls for a more balanced approach to this matter and research on less intrusive technologies for the detection of criminal behaviors online.

Keywords: CSAM, Chatcontrol, Classifiers for CSAM detection, privacy

1. Introduction

Since digital communication became common, the proliferation and spreading of *Child Sexual Abuse Material (CSAM)* became a growing problem that required the intervention of both states and service providers.

In response to this phenomenon, on 11 May 2022, the European Union Commission adopted a legislative proposal to fight the spreading of CSAM and online children's sexual solicitation (grooming)[9]. This proposed regulation is also known as Chatcontrol.

This legislation aims to detect these crimes by applying technologies based on ML classifiers for images and text on intercepted communications on targeted digital platforms.

However, such a proposal raised debates and subsequent alarms on the potential effects that this proposal might have on the privacy of European citizens. In particular, concern was caused by the proposed scanning methodology that would endanger the use of *End-to-End Encryption (E2EE)* [4].

This work explores not only the legislative aspects of this proposal relating to privacy rights in the EU by studying the proportionality of the legislation, but also the technical feasibility, efficiency, and efficacy of the proposed infrastructure in solving the presented problem.

In Section 2 this discussion firstly presents some relevant technical aspects relating to content-scanning technologies and algorithms. Successively, the topic has shifted to

introduce parts of the European legislative framework concerning data and communication privacy. After that, the proposal object of this work is presented.

Moreover, In Section 3 the methodology used for gathering the data is discussed.

Section 4 will show considerations on how the proposed system could raise an alarming number of false positives, how it would perform in an adversarial context, and how detection could be avoided.

Furthermore, an exploration of some judicial cases from the *European Court of Human Rights (ECHR)* is presented to understand how privacy rights, national security, and other social matters are balanced.

Successively, in Section 5 a discussion on the legislation proportionality is given based on the results of the evaluation of the above-cited legislative framework, court rulings, and technical assessments on efficacy and efficiency, with also an opinion on a possible outcome for E2EE technology.

Finally, in Section 6 the conclusion is presented calling for further research in balancing child safety and privacy rights, and on less intrusive technologies for crime detection in digital communications.

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 found that with the lowest threshold, PhotoDNA was able to recognize dissimilar images with a 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 raise low false-positive alerts [20], albeit not in an adversarial environment [18].

2.1.2 ML-based detection

This second method uses *machine learning (ML) 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 technology can detect target behavior 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 behavior 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 fosters higher precision in determining if a user is engaging in the target behaviour [3]. As stated in the state of the art of [3], to obtain the most accurate results the entirety of the conversation needs to be processed.

Regarding previously unknown 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 [17]. In particular, Castrillon-Santana et al. (2018) were able to obtain a 7% error rate in classifying adult/no-adult classification [5], 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% [24].

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)*. In this section 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 Legislative Framework

In the European Economic Area (EEA), the use of these technologies to prevent, detect, and prosecute criminal offenses is strictly regulated. In this section 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" [6].

Nonetheless, Art.52 (3) imposes the same limitations to these rights as the one specified in the *European Convention on Human Rights (ECHR)* [6]. 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 by 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" [10], 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 [13] 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) [13].

Regarding the context of this study, Art. 5 (1) point c of the GDPR [13] 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" [13] (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" [7].

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" [7].

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 [7] [6] [22].

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. The temporary measure of this regulation is because a legislative framework to fight this phenomenon while maintaining 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 [19]. Most important is Art. 3 (1) which prescribes how Art. 5 (1) and 6 (1) of the ePrivacy Directive will not apply anymore concerning the

confidentiality of those communications where there is a well-founded suspicion of online CSA [19].

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) [19]. As per Art. 10, this regulation was set to apply until August 2024 but was extended to apply until the 3rd of April 2026 [19].

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, 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 [9].

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 [9]. Again in Art. 3, the proposal lays down a list of factors helpful in identifying such risks such as previous instances of the target behavior, 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) [9].

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

A qualitative methodology was adopted to conduct this study. In particular, different legal and technical documentation and publications were taken into consideration.

To search the technical material concerning the computer science domain, the following (non-exhaustive) list of keywords was used on Google Scholar and ReasearchGate: classifier for adult children, ML for CSAM detection, PhotoDNA, client-side scanning, and AI classifier for grooming.

Concerning the research related to the juridical aspects, the following keywords were used on EUR-Lex: CSAM, CSAM Regulation, and privacy. Moreover, concerning the search for judiciary cases, the following keywords were used on curia.europa.eu: personal data processing, fundamental rights, ePrivacy, privacy.

Furthermore, to gather additional data on judiciary cases relating to privacy, thematic factsheets for the press released by the ECHR[1] were taken into consideration as indexes for relevant rulings. In particular, factsheets "Mass surveillance" and "Protection of personal data" were considered.

The reason for the review of these documents was to assess the technical efficiency, efficacy, and feasibility of the solution in terms of technology. Moreover, the analysis relating to the juridical aspect was focused partly on the proportionality of the proposal and the interpretation of Article 8 of the Charter of Fundamental Rights of the European Union about both the examined judiciary cases and the CSAM proposal.

4. Findings

4.1 Technical evaluation of the proposal's suitability

The proposal aims 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 previously 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 the client-side before the encryption that happens before the material is sent [8].

While stated in Annex 9, the proposal fails to recognize 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 the hashing algorithm a hash that will collide with CSAM hash[2].

To avoid these types of attacks, the assessment aims at following a security by obscurity design to avoid the leaking of the algorithm. This type of design is not state-of-the-art and is not secure by design by itself [23]. Furthermore, the algorithm will have to be run on the client side, making maintaining the secrecy of the hasher implementation difficult[3].

4.1.2 Detection of unknown CSAM

As previously described, to detect unknown CSAM machine learning classifiers are needed. Discussing the context in which the communication is E2E encrypted, the

[1] see `https://prd-echr.coe.int/web/echr/factsheets`

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

[3] See `https://github.com/AsuharietYgvar/AppleNeuralHash2ONNX`

impact assessment proposes the deployment of the classifier model on the client side and operating the classification previously the encryption of the media. Furthermore, this approach is difficult to implement, especially on smartphones and other devices with lower computational capabilities.

Moreover, both the accompanying and the complementary impact assessment report that there exists a classifier for detecting new CSAM, Thorn's Safer, with a precision of 99.9% and a recall of 80% [8] [12]. Unfortunately, no proof is provided to support such claims and the citation linking to the benchmarking platform redirects to a benchmarking platform for perceptual hashing algorithms and not for classifiers of previously unknown CSAM[4].

Furthermore, the performance reported in these documents is optimistic at least if compared with what has been summarized in section **??**, and also taking into consideration that Thorn's model for CSAM detection is deployed on server side [21] having access to higher computing power.

4.1.3 Sexual solicitation of children

Also, in this case, ML classifiers are required to scan conversations and flag grooming behavior. 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 [8]. Furthermore, server-side-based solutions developed under project Arthemis are reported to have an accuracy of only 88% [8].

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 it on the controlled platform. This issue is also been noted in the EDPB-EDPS Joint Opinion [2] but never 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 a secret is used to compute a symmetric key used to encrypt the media.

The bash commands in listing 4.1.4 show how a user computes both a symmetric and asymmetric key to exchange messages on an unsecured channel, while Figure 1 shows how this method is applied to avoid detection in the chat of a service provider under detection order.

```
openssl ecparam -name prime256v1 -out ecparam.pem

openssl ecparam -name prime256v1 -genkey -noout -out party1_private_key.
    pem

openssl ec -in party1_private_key.pem -pubout -out party1_public_key.pem

openssl pkeyutl -derive -inkey party1_private_key.pem  -peerkey
    party2_public_key.pem -out shared_secret.bin

openssl enc -aes-256-cbc -nosalt -pass file:./shared_secret.bin -in body.
    bin -out body.ecb.bin
```

Listing 1: Client 1 computes the shared secret over an unsecured channel and encrypts an image

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

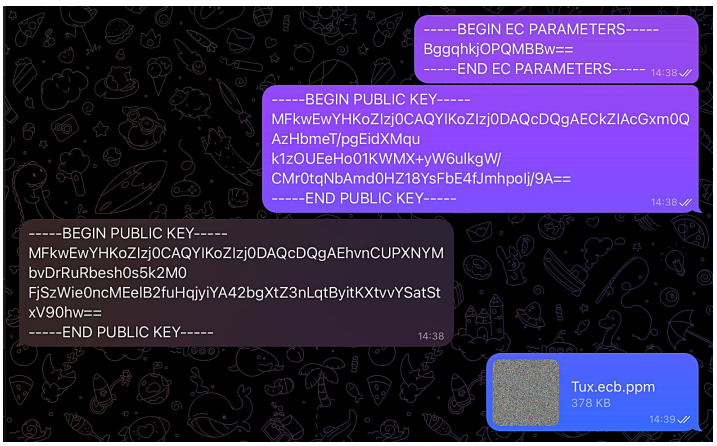

Figure 1: ECDH key exchange and AES encrypted image sharing

4.2 Relevant ECHR and ECJ 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/or 10 of the EU Charter of Fundamental Rights.

4.2.1 ECHR: Big Brother and others v. the United Kingdom

As described in [14], 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 a violation of both Article 8 and Article 10 since secret bulk intervention 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 clearly present the grounds upon which the interception may take place.

4.2.2 ECHR: Breyer v. Germany

As reported in judgment [15], the applicants complained that telecommunication service providers were storing their personal data outside the scope of transactions as prescribed by the Telecommunications Act (Telekommunikationsgesetz), and claimed a violation of Article 8. Such personal data were name, address, phone number, other mobile numbers, date of birth, and contract date.
The Court found that there was no violation of Article 8. More specifically, the Court notes that an interference with the applicant's privacy rights has been taking place as their personal data were being stored by the service provider. Nonetheless, the conclusion was that such interference is justified since the storage modalities, and in relation to the amount of stored data, were sufficiently clear. Furthermore, the Court accepted such interference as the measure was deemed proportional: the aim of fighting terrorism and organized crime is legitimate and the storage of the (limited and well-defined) personal data is proportionate since no sensitive data were gathered and/or processed.

4.2.3 ECHR: Szabó and Vissy v. Hungary

In this judgment, the applicants complained about the disproportionate and unjustified governmental policies for secret surveillance [16]. In particular, the applicants complained about the possibility of the system being abused.
The court found that the scope of the measures could include virtually anyone in the interested country, with the technologies enabling the Government to intercept masses of data easily concerning even persons outside the original range of operation.
For this reason, the Court ruled that a violation of Article 8 was in place.

4.2.4 Grand Chamber: La Quadrature du Net e a.

This judicial case consists of a preliminary ruling regarding, inter alia, the interpretation of Article 15(1) of Directive 2002/58/EC (ePrivacy Directive) previously introduced in section 2.3.3.
Under the cited Article and under the consideration of articles 7, 8, and 11 of the Fundamental Right Charter, in this preliminary ruling, the Court concluded that real-time communication data collection is lawful if "recourse to automated analysis is limited to situations in which a Member State is facing a serious threat to national security which is shown to be genuine and present or foreseeable" [11] and if "recourse to the real-time collection of traffic and location data is limited to persons in respect of whom there is a valid reason to suspect that they are involved in one way or another in terrorist activities and is subject to a prior review carried out either by a court or by an independent administrative body" [11].

5. Discussion

It is clear from what was introduced in section 2.3 that among the foundamental values of the Union, the right to privacy, as expressed in the Charter and elaborated on by various directives and regulations, e.g. ePrivacy, is one of the most needed in a democratic society.
Nonetheless, great discussions are arising in these years to assess to which extends such right can be interfered with in the name of national security and crime prevention and detection.
The proposal that is the object of this research, as part of these discussions, has risen significant concern in light of the interference that it might cause with to the previously defined right to privacy.
The results show both the degree of the technical efficacy and efficiency of the proposed solution, and also report different rulings of the ECHR regarding mass surveillance and the interpretation of the ePrivacy directive and Article 8 of the Charter.

5.1 Relation with current legislative framework and court cases

The proposed regulation imposes service providers under a detection order to process all communications taking place on their platforms and subsequently store the ones for which there is a confirmation of unlawful conduct, i.e. in this case sharing of CSAM material and grooming.
The process laied down in the proposal entails (a) the real-time interception and initial retention of communication data, (b) application of selectors, i.e. ML processing and/or via percpeptual hashing, (c) examination of the alerts by a human analyst, and (d) data retention and reporting.
For this reason, the proposed methodology for detecting such behaviors can be described as the *bulk interception* of communications as in the judgement for case Big Brother and others v. the United Kingdom (see section 4.2.1). In the same court ruling, the Court agreed on viewing these bulk interceptions "as a gradual process in which the degree of

interference with individuals' Article 8 rights increases as the process progresses"[14].
Furthermore, to avoid the implementation of a system of mass surveillance, a proposal that make use of bulk interception should lay down clear and well defined safeguards to avoid the abuse of the system (see Big Brother and others), the collection of unnecesary data, and data gathering of people outside the original range of the operation (see Szabó and Vissy v. Hungary).
It is true that in Szabó and Vissy v. Hungary the risk of system abuse is given by the secretive nature of the operation. Nonetheless, the Chatcontrol proposal makes use of technologies that can be technically and legislatively subverted to detect behaviors different from the original target, e.g. terrorist, political opposition, journalist. In particular, it is possible to tamper with the machine learning model, or injecting hash signatures to raise alerts for different type of behaviours and shared media.
Moreover, it is possible that under the specifications of this proposal there will be interception and storing of communication data belonging to people outside the scope of the operation. Some examples of situations that could lead to such event are (a) communication with sexual tone between two minors[5], (b) communication with sexual tone between two people for which Romeo & Juliet like laws apply, (c) communication with sexual tone in which media of adult that appear young are shared[6].
Furthermore, the interception can be said to be *real-time* since it is applied using both server or client side scanning depending on the communication being E2EE or not. Given this interpretation of the *real-time* adjective, given the preliminary ruling in La Quadrature du Net e a. (see section 4.2.4), and given that the purpose of of the interception is not to fight an imminent or foreseeable terrorist threat, one can argue that such proposal is violating Article 8 of the Charter.
Finally, the conditions for a detection order to be issued to a service provider are generic, e.g. "the extent to which the service is used or is likely to be used by children" and "any previously identified instances of use of its services for the purpose of online child sexual abuse"[9], and the risk assessment lacks strong framework for the computing of the risk like the CVSS framework in the domain of cybersecurity risk assessment. This could lead to a great number of service providers to be issued such orders.

5.2 Efficiency and efficacy

The proposal lay down three technical problems: known CSAM detection, unknown CSAM detection, and grooming detection.
The first problem is solved by the use perceptual hashing. Unfortunaltely, in the proposal there is not any discussion regarding the deployment of such technologies in an adversarial context. As discussed in section 4.1.1, sharing an image in which some adversarial noise was injected can cause an arbitry number of false-positives. Of course, this is possible if there is some knowledge about how the hashing algorithm work. Unfortunately, the algorithm is executed on client side, causing the implementation of a *security by obscurity* framework difficult.
The second problem is tackled by using ML classifiers for which only claims on their efficiency are provided by the proposal impact assessment. In fact, section 2.1.2 reports results that are less optimistic that the ones in the impact assessment concerning Thorn's Safer tool. Furthermore, these models are likely to raise alerts in those situations discussed in section 5.1 above, other than being difficult to deploy in a CSS context given their computational demands.

[5] see statistics on minors that engage in dangerous online behaviors at `https://www-statista-com.ezp.biblio.unitn.it/statistics/1356753/teens-in-europe-engaged-in-selected-types-of-online-behavior/`

[6] Other situations may apply, like `https://www.koffellaw.com/blog/google-ai-technology-flags-dad-who-took-photos-o/`

To solve the third problem, the Commission proposed the use of text-based ML classifiers giving as an example Microsoft tools that have an accuracy of only 88%, in my opinion too low to be implemented in such an extended interception domain.
Furthermore, as explained in section 4.1.4, it is relatively easy to avoid detection on platforms that are subject to detection orders by simply exchange keys as in an unsecure channel.
Finally, it is important to note how difficult it is to implement such detection systems on platform where E2E encrypted communications take place. Not only it is difficult, but it is more costly in terms of time and in an economical sense. As a consequence, there is a possibility that those providers that will be subject to a detection order will choose to remove E2EE functionalities from their platforms altogheter damaging substantilly the privacy of European citizens and the security of their communications on those platforms.

6. Conclusion

This work offered an overview of both the legislative and technical problems related to the European Commission CSAM Regulation Proposal, also known as Chatcontrol 2.0. In particular, to understand the degree of interference that this proposal would cause to the privacy rights of EU citizens, both current related EU regulations and directives were taken into consideration. Furthermore, to interpret some fundamental articles, recent judiciary cases were reviewed.
Moreover, to assess the feasibility, efficacy, and efficiency of the proposed infrastructure from a technical standpoint, research data on classifiers, hashing algorithms, and content-scanning technologies were gathered. In particular, these data were critically compared with the ones reported in the impact assessment of the Commission and assessment of the Parliament.
While this work has limitations in terms of data analyzed and the expertise of the author in terms of legislative and judicial matters, it is my conclusion that this proposition presents an unbalanced solution to the problem of CSAM online spreading and sexual solicitation of children.
This is, not only because the control could easily be bypassed by those who had sufficient reason to learn how to do it, e.g. criminals, but also because the interference with the Fundamental Rights of the Union would not be proportional with the aim, the efficacy, and the efficiency of this proposal.
Ultimately, while researching this matter, not only did it become obvious how some consensus has to be reached on the balance between child safety and privacy, but also it became apparent that more research has to be carried out on non-intrusive (if any exists) or less intrusive technology for crime detection in digital communications.

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] Patrick Breyer. *Chat control: New EU government attempt to bulk search private messages and destroy secure end-to-end encryption*. 2024. URL: https://www.patrick-breyer.de/en/chat-control-new-eu-government-attempt-to-bulk-search-private-messages-and-destroy-secure-end-to-end-encryption/ (visited on 08/17/2024).

[5] 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.

[6] *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).

[7] *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).

[8] 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).

[9] 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).

[10] *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).

[11] Grand Chamber of the European Court of Human Rights. *Judgment of the Court (Grand Chamber) of 6 October 2020. La Quadrature du Net and Others v Premier ministre and Others*. 2020. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:62018CJ0511 (visited on 08/14/2024).

[12] 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).

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

[14] 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).

[15] European Court of Human Rights. *CASE OF BREYER v. GERMANY*. 2020. URL: https://hudoc.echr.coe.int/eng#%7B%22itemid%22:%5B%22001-200442%22%5D%7D (visited on 08/12/2024).

[16] European Court of Human Rights. *CASE OF SZABÓ AND VISSY v. HUNGARY*. 2016. URL: https://hudoc.echr.coe.int/eng#%7B%22itemid%22:%5B%22001-160020%22%5D%7D (visited on 08/12/2024).

[17] 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.

[18] 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.

[19] *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).

[20] 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.

[21] 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).

[22] *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).

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

[24] 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.