Outcome Report of 2nd Expert Workshop on EU proposed Regulation on preventing and combatting online child sexual abuse

Crocker, A.; Leiser, M.R.; Witting, S.K.

SECOND EXPERT WORKSHOP ON EU PROPOSED REGULATION ON PREVENTING AND COMBATTING CHILD SEXUAL ABUSE

jointly organised by the Vrije Universiteit-Amsterdam, ECPAT International and Leiden University

2nd & 3rd March 2023, VU – Amsterdam

OUTCOME REPORT

Prepared by:

Amy Crocker, Head of Child Protection and Technology, ECPAT International

Dr Mark R. Leiser, Assistant Professor for Digital, Internet Law, and Platform Regulation at VU-Amsterdam

Dr Sabine K. Witting, Assistant Professor for Law & Digital Technologies at Leiden University

ACKNOWLEDGEMENTS

This Expert Workshop on the EU Proposed Regulation on Preventing and Combatting Child Sexual Abuse was jointly organised and funded by Vrije University-Amsterdam, ECPAT International and Leiden University. Thanks to Arno Lodder at Vrije University-Amsterdam Law and Technology Institute for all the support. The organisers also thank Bente Eelman from Vrije University-Amsterdam and Iona Cable from ECPAT International for notetaking throughout the workshop.

DISCLAIMER

Please note that this report reflects the broad scope of presentations and discussions during the workshop. Therefore, the report's content does not necessarily reflect the full view of the organisers' from Vrije University-Amsterdam, ECPAT International and Leiden University.

Furthermore, please note that the content of the report combines the content of presentations and discussion at the workshop with additional context information. The views and opinions referred to in the report are conveyed in good faith as accurate.

The authors acknowledge that the use of acronyms to describe child sexual abuse is common practice but does carry the risk of abstracting the reader from the reality of this severe harm to children.

ABBREVIATIONS

API	Application Programming Interface
CSA	Child Sexual Abuse
CSAM	Child Sexual Abuse Material
CSS	Client-side scanning
EDPB	European Data Protection Board
EC	European Commission
EU	European Union
E2EE	End-to-end encryption
EUROPOL	European Union Agency for Law Enforcement Cooperation
MD5	Message Digest Algorithm 5
P2P	Peer-to-Peer
SHA-1	Secure Hash Algorithm 1
TC	Technology Committee

DEFINITIONS

The definitions below are taken from the 2022 EU proposed Regulation laying down rules to prevent and combat child sexual abuse[1] (hereafter 'proposed Regulation'). Any articles cited in the below definitions are those of the proposed Regulation. Directive 2011/93/EU refers to the Directive on combating the sexual abuse and sexual exploitation of children and child pornography (hereafter 'CSA Directive')[2]:

Child	means any natural person below the age of 18 years.
Child sexual abuse material	means material constituting child pornography or pornographic performance as defined in Article 2, points (c) and (e) of Directive 2011/93/EU.
Child sexual abuse offences	means offences defined in Articles 3 to 7 of Directive 2011/93/EU.
Known child sexual abuse material	means potential child sexual abuse material detected using the indicators in the database of indicators referred to in Article 44(1), point (a) of the proposed Regulation.
New child sexual abuse material	means potential child sexual abuse material detected using the indicators in the database of indicators referred to in Article 44(1), point (b) of the proposed Regulation.
Online child sexual abuse	means the online dissemination of child sexual abuse material and the solicitation of children.
Solicitation of children	means soliciting children for sexual purposes as referred to in Article 6 of Directive 2011/93/EU.

[1] Proposal for a Regulation of the European Parliament and of the Council laying down rules to prevent and combat child sexual abuse, COM(2022) 209 final, 11.5.2022.
[2] Directive 2011/93/EU of the European Parliament and the Council of 13 December 2011 on combating the sexual abuse and sexual exploitation of children and child pornography, OJ L 335, 17.12.2011.

TABLE OF CONTENTS

1. BACKGROUND TO THE WORKSHOP SERIES

The proposed Regulation by the European Commission (EC) laying down rules to prevent and combat child sexual abuse[3] (hereafter 'proposed Regulation') has sparked lively debate. The overall objectives of the proposed Regulation have been broadly welcomed by child protection organisations, arguing that voluntary action including detection measures have proven insufficient to combat the evolving nature of online child sexual abuse (CSA). They call on the European Union (EU) to stricter regulate the role of online platforms in preventing and responding to online CSA, not only to prevent the further dissemination of child sexual abuse material (CSAM), but also to detect, disrupt and report the solicitation of children for sexual purposes, and to support efforts to rescue children from ongoing abuse.

While the goal of protecting children is a common one, concerns have been voiced from a range of groups that mandatory detection measures, if imposed at scale under the proposed Regulation, would violate the rights to data protection, privacy, and free expression as set out under the Charter of Fundamental Rights of the European Union (hereafter 'EU Charter')[4], particularly if applied in the context of end-to-end encryption (E2EE). While child protection organisations in their majority are vocal in support of E2EE as a vital security measure to protect sensitive information from unauthorised access, they call for technical solutions that prevent E2EE from providing safe havens for those who abuse and exploit children. Further concerns are that the proposed Regulation might lead to the general and indiscriminate scanning of communication data and that automated detection technologies could lead to false positives, resulting in legitimate content being flagged and removed. This could have implications for free speech and potentially harm businesses relying on online platforms to reach customers.

The debate over the proposed Regulation highlights the challenge of balancing privacy and security with children's right to protection from violence and the existing EU obligation to protect children from violence, abuse, and exploitation in a digital age. While E2EE is essential for protecting sensitive information, it is also essential that action is taken to combat illegal online activities such as online CSA. The challenge for policymakers and industry leaders is finding a way to achieve both objectives without compromising one or the other. As the proposed Regulation continues to be debated, it will be essential to understand the facts of the proposed technological solutions and processes and to consider the potential implications for E2EE and other digital security measures.

Against this background, the second in a series of workshops was held at Vrije Universiteit-Amsterdam on 2nd & 3rd March 2023 to facilitate open discussions and identify areas of common ground with technical experts from various fields relevant to the proposed Regulation, including child rights, privacy, data protection, platform regulation, and fundamental rights. The first workshop at Leiden University in October 2022 discussed the existing EU legal framework relevant to the proposed Regulation and its impact on fundamental rights under the EU Charter. After the first workshop in Leiden and in consultation with ECPAT International and the Council of Europe's Lanzarote Committee

[3] Proposal for a Regulation of the European Parliament and of the Council laying down rules to prevent and combat child sexual abuse, COM(2022) 209 final, 11.5.2022.

[4] European Union: Council of the European Union, Charter of Fundamental Rights of the European Union (2007/C 303/01), 14 December 2007, C 303/1.

Secretariat, Dr Mark Leiser and Dr Sabine Witting produced an outcome report that was presented to the Council of Europe's Lanzarote Committee in Strasbourg in February 2023.

Based on the interests of the participants at the first workshop at Leiden University and reflecting a central focus of the debate around the proposed Regulation, a further expert workshop was scheduled to discuss the technology-related aspects of the proposed Regulation. Key themes, including E2EE and detection technologies, were discussed to support a constructive conversation about whether and how the proposed Regulation in its current form can be applied in the existing legal and technological landscape. The first day of the workshop focused on discussing the technical implications of detection technologies and E2EE. The main objective was to help all participants understand the relevant provisions in the proposed Regulation, the current state of the art of detection technologies, the functioning and purpose of E2EE, and the associated risks related to fundamental rights, including privacy and security. Day 2 of the workshop aimed to reassess the rating of the technological solutions set out on Day 1 across various criteria, focusing on privacy and security while considering the value of their application for the purposes of child protection and crime prevention. Additionally, it aimed to take a distinct fundamental rights perspective to assess which risks are imposed on fundamental rights and what procedural/substantive safeguards might be required in the proposed Regulation to protect such rights.

The outcome of the discussions led to recommendations set out at the end of this report (see section 5). These recommendations should be used to discuss potential amendments to the relevant provisions dealing with technology and related safeguards outlined in the proposed Regulation.

2. THE ROLE OF TECHNOLOGY IN COMBATING CHILD SEXUAL ABUSE ONLINE

2.1 SCALE, SCOPE, AND COMPLEXITY OF THE PROBLEM

Online child sexual abuse represents a significant, complex, and egregious dimension of the violence, abuse, and exploitation that children worldwide suffer each day. The scale, scope, and complexity of the problem of online CSA are vast. While challenging to understand purely through the filter of numbers, example survey data shows that up to 20% of children had experienced online sexual abuse or exploitation in the past 12 months (Disrupting Harm, 2022[5]) and that 54% of young people (57% of all girls and 48% of all boys) reported experiencing online sexual harm before they were 18 (WeProtect Global Alliance, 2022[6]). According to the EU Kids Online report for 2021, between 21% (France) and 50% (Serbia) of 9- to 16-year-olds reported seeing some sexual image in the past year.

In terms of CSAM and solicitation of children, the Internet Watch Foundation's Annual Report for 2021 showed that the organisation took action to remove 252,000 URLs with images or videos of children being raped and suffering sexual abuse. Of those URLs, 182,281 contained images or videos of "self-generated" material and sexual abuse imagery of children aged 11-13 was most prevalent. In 2021,

[5] https://www.end-violence.org/disrupting-harm

[6] *A global study of childhood experiences of 18–20-year-olds*, Economist Impact, 2022, available at: https://www.weprotect.org/economist-impact-global-survey/#report

the NCMEC CyberTipline received 29.1 million reports of exploitation and abuse, most coming from online service providers complying with their obligation to report child sexual abuse on their platforms once made aware of its presence.

2.2 PREVENTION AND RESPONSE

Most online child sexual exploitation and abuse are not disclosed to anyone or reported to a helpline or the police. This is due to a range of factors, including that many victims are unaware that their experience constitutes abuse, or are too young to speak let alone make a report. For other victims, feelings of shame and fear linked to sexuality, social taboos and gender norms in society are frequently compounded by the solicitation process and contribute to a culture of silence. And while so called “self-produced" content is increasingly being seen circulating online, distinguishing content that has been voluntarily and consensually produced from content that has been coerced is not possible from visual analysis alone. Regardless of any action that will be taken in response to the content, its removal from wide online distribution is essential. For victims of abuse depicted in CSAM, the knowledge of sexual abuse images and videos circulating online can have a severe and long-term impact. Furthermore, significant societal and economic costs are associated with violence against children.

While an optimal response to online child sexual exploitation and abuse requires a range of human and technological measures, cooperation between online service providers, hotlines, and law enforcement must continue to evolve and be met with resources, noting that existing capacity and resource gaps in law enforcement cannot be resolved overnight. While practitioner views offer insight into the effort needed to respond to online child sexual exploitation and abuse, public opinion plays a crucial role given the implications of child abuse and technology in general on society, in the context of the regulation. In a survey of 9,410 adults from France, Germany, Hungary, Italy, the Netherlands, Poland, Sweden, and Spain, conducted by Defence for Children - ECPAT Netherlands and ECPAT International in 2021, over 75% of people across the eight countries believed that protecting children from abuse was as important or more important than the protection of their privacy. Furthermore, 68% supported a legal requirement for online service providers, such as social media platforms, to use automated technology tools to detect and flag signs of online child sexual exploitation and abuse.

2.3 THE ROLE OF TECHNOLOGY

The role of technology in addressing the issue of online child sexual abuse is significant. Detection technology can triage high volumes of content, detect "known" CSAM, flag high-risk "new", i.e., unknown content, flag high-risk written exchanges, and support content moderation. However, human moderation is also essential. People can moderate activity on open platforms, validate items flagged by technology, and act upon and escalate the response to content, conduct, and contact that poses a risk to a child.

In the context of this workshop, some participants are firm in support for the use of detection technologies for preventing and combating child sexual abuse online, calling for a risk assessment-based safety-by-design approach and strong safeguards around detection technologies to prevent misuse. Another participant emphasised the importance of maintaining an ongoing legal basis for the voluntary use of detection technologies and establishing an EU Centre with a strong victim support mandate.

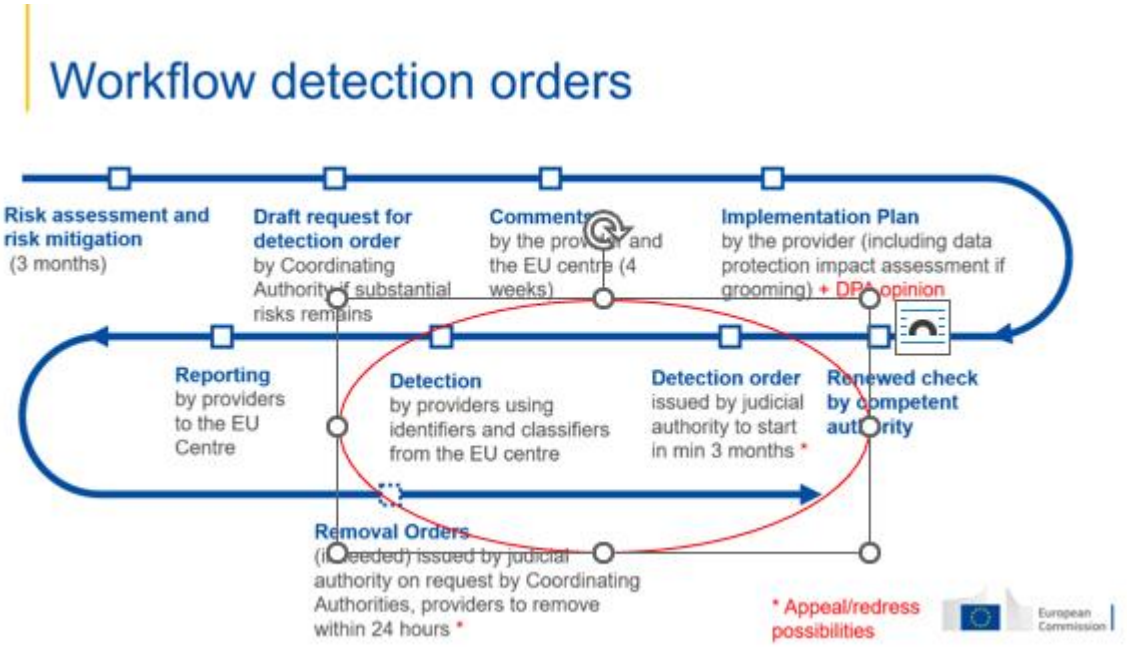
2.4 CONCLUSION

The proposed Regulation is a step towards addressing the issue of online CSA. The scale, scope, and complexity of the problem are vast, and improved cooperation is needed between online service providers, hotlines, and law enforcement. The role of technology in detecting and flagging online CSA is crucial. Still, human moderation is also essential. Strong safeguards before the deployment of detection technologies provide a useful framework for ensuring that the technology is used to protect children's rights, preserve privacy and confidentiality, and prevent misuse. The proposed Regulation's legal and policy context underscores the issue's significance and highlights the importance of addressing it from a legal standpoint. The human rights implications of technology deployment are essential society considerations. Using technology that complies with the EU Charter obligations can work towards preventing and combatting online CSA and ensure that children are protected from the devastating effects of online exploitation and abuse.

3. RELEVANT LEGAL PROVISIONS IN THE PROPOSED REGULATION CONCERNING DETECTION TECHNOLOGIES

This section briefly highlights the relevant provisions in the proposed Regulation dealing with detection technologies, including their potential deployment in E2EE environments. These provisions are the current 'baseline' in terms of procedural safeguards set forth in the proposed Regulation. Recommendations on how these provisions can be improved to prevent the violation of fundamental rights will be discussed in section 5 (Recommendations).

As a reminder, below is a summary of the workflow for issuing a detection order under the proposed Regulation.



Based on the EC's workflow, the Workshop operated on the hypothetical assumption that a valid detection order has been sent to a platform with interpersonal messaging services embedded in its services. Because the first Workshop had explained this particular procedure at length, the remainder of the second workshop focused on the implementation of the protection order.

3.1 ART 7 (1) IN CONJUNCTION WITH ART 7 (8): ISSUING A DETECTION ORDER FOR E2EE PLATFORMS

Some argue that the discussion around the impact of the proposed Regulation on E2EE is misguided as Coordinating Authorities would not issue a detection order for an E2EE platforms in the first place as this would be considered disproportionate in the context of Art 7 (8)[7], second paragraph:

> *'To that aim, [the Coordinating Authorities] shall take into account **all relevant parameters, including the availability of sufficiently reliable detection technologies** in that they limit to the maximum extent possible the rate of errors regarding the detection and **their suitability and effectiveness for achieving the objectives** of this Regulation, as well as the impact of the measures on the rights of the users affected, and require the taking of the **least intrusive measures**, in accordance with Article 10, from among several equally effective measures.' [Emphasis added]*

Using statutory interpretation techniques[8], such an interpretation might be possible but is by no means obvious from the wording of the provision. Hence it cannot be argued that Art 7 (8) is a sufficient safeguard to prevent the issuing of detection order for E2EE platforms. This leads us to the question of what happens once a platform has been issued with a detection order (see Art 10 below).

3.2 ART 10 (1): EXECUTING A DETECTION ORDER

Art 10 (1) is the relevant provision setting out the legal obligations for a platform once subjected to a detection order:

> *'Providers of hosting services and providers of interpersonal communication services that have received a detection order **shall execute it by installing and operating technologies to detect the dissemination of known or new CSAM or the solicitation of children**, as applicable, **using the corresponding indicators** provided by the EU Centre in accordance with Article 46.' [Emphasis added]*

The literal interpretation of Article 10 raises several questions, including the meaning of "execute it", "installing and operating technologies," and the use of indicators provided by the EU Centre. The Workshop operated on the assumption that this provision creates a legal obligation on the platform to respond to the detection order and must install and operate technologies to detect the dissemination of known, new CSAM and solicitation, as specified in the detection order.

3.3 ART 10 (3): CRITERIA FOR DETECTION TECHNOLOGIES

[7] Unless referenced otherwise, all articles in this report are those of the Proposal for a Regulation of the European Parliament and of the Council laying down rules to prevent and combat child sexual abuse, COM(2022) 209 final, 11.5.2022.

[8] Statutory interpretation is a crucial aspect of legal analysis that involves looking at the meaning of the law and its application. There are different approaches to statutory interpretation, including the literal, purposive, and teleological approaches. The literal approach involves giving the words their ordinary meaning, while the purposive approach involves looking at the general purpose of an act. The teleological approach is based on the purpose, direction, or design of the text/legislation.

Art 10 (3) sets out the criteria relevant for the selection of suitable detection technologies:

> *'The technologies shall be:*
>
> *(a)* ***effective*** *in detecting the dissemination of known or new child sexual abuse material or the solicitation of children, as applicable.*
>
> *(b)* ***not be able to extract any other information*** *from the relevant communications* ***than the information strictly necessary*** *to detect, using the indicators referred to in paragraph 1 […];*
>
> *(c)* ***in accordance with the state of the art*** *in the industry and the least intrusive in terms of the impact on the users' rights to private and family life, including the confidentiality of communication, and to protection of personal data.*
>
> *(d)* ***sufficiently reliable****, in that they limit to the maximum extent possible the rate of errors regarding the detection.' [Emphasis added]*

These criteria function as safeguards to avoid the deployment of detection technologies which violate fundamental rights. Participants were largely of the opinion that the criteria set forth by the proposed Regulation are broad and open to interpretation. While it is acknowledged that legislation cannot be too specific as it needs to apply to a wide range of technologies, it is not possible to foresee how Art 10 (3) will be interpreted and which type of detection technologies will meet the criteria set forth under Art 10 (3).

3.4 ART 10 (4): RESPONSIBILITIES OF THE PROVIDER WHEN DEPLOYING DETECTION TECHNOLOGIES

Art 10 (4) puts additional responsibilities on the provider when deploying detection technologies:

> *'The provider shall:*
>
> *(a) take all the necessary measures to ensure that the technologies and indicators, as well as the processing of personal data and other data in connection thereto, are used for the* ***sole purpose of detecting the dissemination of known or new child sexual abuse material or the solicitation of children, as applicable,*** *insofar as strictly necessary to execute the detection orders addressed to them.*
>
> *(b)* ***establish effective internal procedures to prevent and, where necessary, detect and remedy any misuse of the technologies****, indicators and personal data and other data referred to in point (a), including unauthorized access to, and unauthorised transfers of, such personal data and other data.*
>
> *(c)* ***ensure regular human oversight*** *as necessary to ensure that the technologies operate in a sufficiently reliable manner and, where necessary, in particular when potential errors and potential solicitation of children are detected, human intervention.*
>
> *(d)* ***establish and operate an accessible, age-appropriate and user-friendly mechanism*** *that allows users to submit to it, within a reasonable timeframe,* ***complaints about alleged infringements*** *of its obligations under this Section […].' [Emphasis added]*

These provisions provide important safeguards to counter some of the risks associated with detection technologies, i.e., the risk of repurposing (Art 10 (4) (a)) or the risk of compromising of detection technologies (Art 10 (4) (b)). Whether these safeguards are sufficient in the context of the proposed detection technologies will be discussed in section 5 (Recommendations).

3.5: ART 50 (1): PROPOSING SUITABLE DETECTION TECHNOLOGIES

Art 50 (1) describes the procedure the EU Centre must follow when making available certain detection technologies:

> *'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).***
>
> *To that aim, the EU Centre shall **compile lists of such technologies**, having regard to the requirements of this Regulation and **in particular those of Article 10(2).***
>
> *Before including specific technologies on those lists, the EU Centre shall **request the opinion of its Technology Committee and of the European Data Protection Board**. […]' [Emphasis added]*

When deciding whether a particular detection technology can be deployed by providers for the execution of detection orders, the EU Centre needs follow the requirements set out in this provision, in particular Art 10 (2). While it is assumed that this cross-reference is aimed to highlight Art 10 (3), the EU Centre shall seek the opinion of the so-called technology committee (TC) and the European Data Protection Board (EDPB). Using statutory interpretation, this provision means that the opinions of these two bodies only need to be sought, but do not have to be considered in the decision-making process.

3.6 UPCOMING GUIDELINES (ART 11) AND RECITAL 26

According to Art 11, the EC may issue guidelines on the application of Articles 7 to 10, having due regard to relevant technological developments and the manners in which the services covered by those provisions are offered and used. Such secondary regulation will hence be crucial to provide more detail to the interpretation of the provisions on issuing and implementing detection orders. However, it must be noted that such Guidelines are not legally binding and can hence not be used to create additional legal safeguards. If additional legal safeguards are required, these need to be incorporated into the actual legal text of the proposed Regulation.

Recital 26 plays an important role for the subject matter of this report as it is the only recital which directly speaks about E2EE. Similarly, it is important to note that Recitals are not legally binding and hence are mere interpretative tools for the substantive provisions of the proposed Regulation.

4. REASSESSING THE VIABILITY OF DETECTION TECHNOLOGIES AND THEIR IMPACT ON FUNDAMENTAL RIGHTS

In this workshop session, participants reassessed the impact of detection technologies on fundamental rights across various criteria. This step was crucial to understand the different levels of impact across different technologies within open communications and E2EE communications. The baseline for this assessment was the rating of the detection technologies in the EC's 2022 Impact

Assessment report[9] (Annexes 8 and 9) which included the criteria of effectiveness, feasibility, privacy, security, and transparency. Participants were invited to consider and reassess the ratings of the EC through a lens of fundamental rights. Based on this assessment, participants went on to propose legal safeguards for the proposed Regulation (section 5), following the rule of thumb that 'the higher the risk, the more stringent the legal safeguard'.

The following detection technologies were selected for deeper analysis and discussion, as these are the detection technologies which the European Commission's Impact Assessment report (Annexes 8 and 9) considers currently deployable or likely to be ready for deployment soon:

Open communications:

1. Hashing technologies
2. Classifiers and AI
3. Text-based communications analysis tools

E2EE communications:

4. On-device full hashing with matching at the server
5. On-device partial hashing with remaining hashing and matching at the server
6. Secure enclaves in the ESP server

Assessment criteria:

Participants applied the below criteria to assess each of the six detection technologies mentioned above:

- Type of online CSA which can/cannot be detected.
- Method of compliance with detection order.
- Impact on Confidentiality of Communications.
- Impact on Security of Communications.
- Risk of Repurposing.

Participants took a distinct fundamental rights perspective to assess the risks imposed on fundamental rights and the procedural/substantive safeguards required for each solution.[10] The following guiding questions were offered to participants:

1. What are the opportunities, pitfalls, and hazards of each type of detection technology mentioned explicitly in the proposed Regulation?

2. What are the risks and gains for fundamental rights associated with their deployment? Does the specific risk amount to a fundamental rights interference or a violation?

[9] EU Commission, *Impact Assessment Report accompanying the Proposal for a Regulation of the European Parliament and of the Council laying down rules to prevent and combat child sexual abuse, SWD(2022), 209 final*, 2022.

[10] Note: organizers purposely skipped over homomorphic encryption due the European Commission's admission that this technology was not close to widespread deployment, see EU Impact Assessment Report on p. 307.

3. On what grounds can the interference be justified? Can the interference be 'cured' or remedied with legal substantive and/or procedural safeguards relative to the amount of interference or risk of violation?

The below sections summarise the discussions amongst participants on the reassessment of the six detection technologies across the proposed criteria, using the guiding questions above to take a fundamental rights perspective.

4.1 HASHING TECHNOLOGIES (OPEN COMMUNICATIONS)

Disclaimer: This section summarises the discussions amongst participants. The authors take responsibility for any inaccuracies in interpretation but are not responsible for the accuracy of the statements made during the workshop.

Hashing is a technique that creates a unique digital fingerprint of a file, such as an image or a video. This fingerprint, a hash value, is a fixed-length string of characters generated using a mathematical algorithm. The hash value is unique to the specific file and cannot be used to reconstruct the original content. When a competent authority receives a new file, it can generate the hash value and compare it to a database of known CSAM hashes. If there is a match, the platform can take action to remove the file and/or report it to the relevant authorities.

Type of online CSA which can/cannot be detected:

Hashing technologies can in general only detect known CSAM, meaning CSAM that is already hashed and contained in a list that is verified by law enforcement to contain CSAM as defined in the relevant jurisdiction. There are two different types of hashing currently used, cryptographic hashing and perceptual hashing. Cryptographic hashing can only detect the exact same imagery, hence cannot detect even slightly modified known CSAM. In contrast, perceptual hashing can detect modified known CSAM to a certain extent (minor content-preserving modifications).

Hashing cannot be deployed to detect unknown imagery, livestreamed abuse, or text-based solicitation.

Method of compliance with detection order:

The platform needs to be connected to a database of verified hashes. No human review is required as the hashes have already been assessed before they were added to the database (see limitations below).

The deployment and effectiveness of hashing technology is influenced by the size and quality of the databases used. If the database contains incomplete or inaccurate data, it may not detect all known CSAM. Moreover, significant obstacles to data sharing and access have necessitated the development of multiple databases that are used by different organisations, that not all stakeholders have access to, across sectors. This means that platforms detect and remove CSAM based on the hash databases to which they have access.

Impact on confidentiality of communications:

Some participants argued that the risk to confidentiality is low: The digital fingerprint generated by the hashing process is just a series of numbers, and the hashing process is essentially irreversible.

However, participants also noted that some hashing techniques are considerably more robust than others, which has implications for the impact of hash detection technologies on the confidentiality of communications. For example, many organisations are moving away from using hashing types such as MD5 and SHA-1, both cryptographic hash types, as these have been shown to have weaknesses and can potentially be reversed. Well established forms of perceptual hashing, developed by technology platforms and experts, also exist, are irreversible and have been in use for many years for the detection of CSAM at scale.

Impact on security of communications:

The use of hashing detection technologies raises concerns about the impact on the security of communications. There are two main concerns: the potential for false positives and the potential for abuse of the technology.

One of the main concerns about the use of hash detection technologies is the potential for false positives. False positives occur when a legitimate file is flagged as CSAM, for example because its hash has been mistakenly included in a CSAM hash database (the file has no relevance whatsoever to the crime of child sexual abuse). It has to be noted that some participants stated that this is extremely rare. The existence of different hash databases that are carefully categorised according to legal and/or other criteria (such as severity of the depicted abuse) is also a function of the fact that legal definitions of CSAM vary between jurisdictions. This is one key reason for human review, which can determine if a particular file is illegal in the jurisdiction in which action (such as removal) against that file is being requested. In some jurisdictions, illegality is not determined by law enforcement but by the courts, and classifications may vary between jurisdictions. Therefore, an appeal and review process are necessary to ensure that a person can challenge the categorisation of an original hashed imagery as illegal.

Risk of repurposing:

To determine the risk of repurposing, it is important to differentiate between the risk of a CSAM hash database being compromised and the risk of legally extending the use of hashing technologies to other illegal or harmful content. It was noted that compromising the database, e.g., by adding non-CSAM imagery to the database and then flagging and removing legitimate content, was theoretically possible. Therefore, ensuring that the data sources used in the technologies are not compromised or tampered with is essential. Transparency reports and random audits can provide an accurate account of how the data sources were obtained, verified, and used in the technology, thereby ensuring the integrity of the data sources.

It was noted that hashing technology is already multi-purpose and deployed in a vast range of public and commercial services used by people every day, all over the world. This means the legislator could expand the mandate for using hashing technology to other illegal or harmful content. This risk can only be mitigated by a strong statement in the proposed Regulation that the legislator will only propose detection technologies, such as hashing, to detect CSAM.

4.2 CLASSIFIERS FOR IMAGERY (OPEN COMMUNICATIONS)

Disclaimer: This section summarises the discussions amongst participants. The authors take responsibility for any inaccuracies in interpretation but are not responsible for the accuracy of the statements made during the workshop.

Classifiers are a form of artificial intelligence (AI) that sorts data into labelled classes or categories. The AI can be trained to detect specifics of image contents, e.g., child nudity or sexual activity involving a child. The AI then learns to distinguish the contents of images based on specific criteria. The training of the AI needs to be based on a significant data set.

Type of online CSA which can/cannot be detected:

A classifier can accurately detect known and unknown CSAM imagery. Detection of online CSA live streaming is theoretically possible, but currently still a technical a challenge for classifiers, as the content is not stored and can only be detected in real-time.

Method of compliance with detection order:

While content classifiers can detect and flag new CSAM with high accuracy, human moderators still need to review all imagery to determine the accuracy of the results. Human moderation remains an essential aspect of using classifiers for detecting CSAM. One of the significant risks involved in using classifiers for detecting CSAM can be mitigated with access to accredited technologies, with clearly defined criteria for accreditation based on EU or other frameworks and standards for the regulation of AI. It is therefore important to ensure that the technology used for detecting CSAM is accredited by regulatory bodies to ensure its reliability and accuracy, among other criteria.

To reduce the amount of imagery requiring human review, good practice by organisations working on the detection, reporting, and removal of CSAM is to take a staggered approach to using technologies, i.e., to first compare all imagery to the hash database using hash-matching technology, and then run classifiers on the remaining imagery to triage potential CSAM from other, non-pertinent content. Human review is for verification of triage results. Even if such a staggered approach is taken, a considerable amount of human review is required. In the context of the proposed Regulation, this makes outsourcing by companies to third parties in and outside the EU very likely, in turn requiring minimum standards and safeguards around workforce training, conditions and wellbeing of content moderators.

Impact on confidentiality of communications:

The impact on confidentiality of communications was considered medium, considering that all content flagged as potential CSAM requires human review. There is also a risk that the awareness of communications being scanned will cause a chilling effect on users, leading to self-censorship and a reduction in the free flow of information.

Another concern with using classifiers for detecting CSAM is outsourcing personal data and confidential communications. Considering the enormous amount of imagery requiring human review, it is likely that platforms outsource this task to non-EU countries to minimise costs. Online platforms may need to share data with third-party service providers to develop and implement classifiers. This

raises questions about protecting personal data and the confidentiality of communications. There may be risks of data breaches, as well as concerns about the misuse of personal data by third-party service providers.

One of the main concerns with using classifiers to detect CSAM is the potential for disproportionate use. It is essential to assess the level of intrusiveness of using classifiers to detect CSAM and to balance this with the protection of children. While using classifiers may be necessary to identify potentially illegal content, ensuring that this does not come at the cost of violating individuals' general privacy and rights is crucial. It is essential to balance the methods used to protect children and the right to privacy and freedom of expression.

Impact on security of communications:

Firstly, it is essential to note that using classifiers for images and videos does not involve any manipulation or compromise of communication. Classifiers are designed to assess and identify specific features of an image or video, such as the presence of CSAM. This means that the use of classifiers does not involve accessing or modifying any communications that are passed through the system. Furthermore, trained professionals typically use classifiers subject to strict protocols and regulations. These professionals must adhere to specific standards and procedures to ensure the classifiers are used appropriately and securely. This includes ensuring that the classifiers are used in a secure environment and that any data collected is handled securely and confidentially.

Risk of repurposing:

Repurposing would occur if classifiers were used for unintended purposes, negatively affecting the tool's reliability and effectiveness. The risk of malicious repurposing of classifiers is considered quite low since it would require significant resources and expertise to modify the tool for unintended purposes as the tools require significant training on large datasets, and every step in the process would need to be repurposed. It is impossible to *accidentally* repurpose a classifier, as doing so would require a deliberate effort and determination to modify the entire training data set or algorithm.

One factor that could increase the risk of repurposing is the bias in the population of data sets and the deployment of biased algorithms on those datasets. Human intervention is required to train classifiers and moderate their outputs, which can enable attackers to exploit weaknesses. For example, a malicious actor could attempt to manipulate the training data to bias the classifier towards identifying certain types of content over others. However, such attacks are unlikely to succeed in practice, as human moderators are typically highly trained and vigilant, and their work is subject to rigorous quality control procedures.

Despite the risk of repurposing through compromising the initial data set, there is also a risk of legal repurposing of such a system, i.e., the legislator extending the use of such detection technology to other areas of illegal or harmful content. Such legal repurposing is a real risk which can only partially be mitigated, e.g., by adding a clause to the proposed Regulation which prohibits such legal repurposing. However, there is always the chance that such a clause is overridden by subsequent regulation. It must be noted, however, that this risk exists for a wide range of technology deployed in our daily lives and is not specific to the issue of online CSA detection technologies.

In conclusion, the risk of repurposing classifiers for detecting CSAM is significant. Still, it is relatively low compared to the benefits of using such tools to combat this form of abuse. While intentional repurposing could have negative consequences on the reliability and effectiveness of classifiers, the low risk of such actions being successful, combined with the vigilance of human moderators, means that the benefits of using classifiers outweigh the risks. However, developers and users of these tools must remain vigilant and continue to monitor and mitigate the risk of repurposing as these technologies evolve and new threats emerge.

4.3 CLASSIFIERS FOR TEXT (OPEN COMMUNICATIONS)

Disclaimer: This section summarises the discussions amongst participants. The authors take responsibility for any inaccuracies in interpretation but are not responsible for the accuracy of the statements made during the workshop.

The implementation of classifiers for text-based analysis is like that of classifiers for image analysis. In both cases, a database is needed, and classifiers are used to identify patterns in the content.

Type of online CSA which can/cannot be detected:

Text-based classifiers can detect conversations indicative of solicitation. It must be noted that existing text-based classifiers are trained on English language models and hence not yet available for any other languages, which limits their effectiveness. The EU has 24 official languages, and there may be differences in how solicitation is expressed in different languages. This means that it is vital to have linguistically diverse teams of analysts, such as those available at hotlines and within dedicated teams at online service providers who can interpret the content of the conversation accurately. Machine learning algorithms may also be helpful in this context, as they can be trained to recognise patterns of language use across different languages.

Method of compliance with detection order:

To detect and prevent such activity, different methods of text-based analysis have been proposed. One of the key issues in implementing classifiers for text-based analysis is the accuracy of language scanning. While text-based analysis has the potential to identify patterns and identify solicitation language, it is far more dependent on contextual factors, including metadata and semantics, than image analysis. This would require the collection and analysis of more communications data and/or metadata.

The language in the early stages of solicitation might be more prone to detection by a text-based classifier because an offender must achieve a specific goal, and the victim may have less linguistic flexibility. This suggests that early detection of solicitation may be possible through text-based analysis. At the same time, perpetrators often use coded language to avoid detection.

Participants highlighted the need for a considerable number of professionals to carry out the human review of flagged communications. Further, language use varies across cultures, and what may be considered normal or acceptable in one culture may not be the same in another. Therefore, classifiers must be developed with a deep understanding of the cultures and languages they monitor. Failure to

do so may result in biases that may result in innocent communications being flagged as potential cases of solicitation.

Impact on confidentiality of communications:

General monitoring of all communications is controversial, as it can infringe on the right to privacy and freedom of expression. However, it may also be necessary in some cases to ensure that solicitation is detected and prevented. In such cases, it is important to have a robust human review system to ensure that only genuinely suspicious communications are investigated further. This can help reduce the risk of false positives and minimise the impact on individual privacy and freedom of expression.

One possible solution to use text-based classifiers minimally invasively is to adopt a phased or multilevel approach to the communications analysis. In the first instance, only suspicious keywords or phrases are flagged for further investigation. This would help reduce the risk of false positives and minimise the impact on the individual's privacy. Building on the initial analysis, the next phase could include analysing the content of the conversation in more detail, looking for patterns of behaviour or using machine learning algorithms to identify common characteristics of solicitation. By building on the initial analysis results, it would be possible to develop a more comprehensive picture of the conversation without compromising communications confidentiality.

Another possible approach is to use a fragmented hash approach to analyse communications. This involves breaking the conversation into small fragments; each assigned a unique hash value. The hash values are then compared to a database of known solicitation techniques to identify potential matches. This approach has the advantage of reducing the amount of data that needs to be analysed, making it more efficient and reducing the risk of false positives. However, it may also be more invasive than other approaches, as it involves analysing the content of the conversation in more detail.

One of the biggest challenges in deploying text-based classifiers is the need for more contextual information. To achieve accurate results, classifiers require a large set of contextual information, which can be challenging. Without the necessary contextual information, classifiers may struggle distinguishing between solicitation and innocent conversations, resulting in high false positive rates. Using metadata to triage suspected solicitation is often proposed as less privacy-intrusive measure. However, metadata can reveal sensitive information about individuals' communication patterns, including their contacts, locations, and online activities. The use of metadata in complementing classifiers must be approached with caution, and appropriate safeguards must be put in place to protect individuals' privacy.

The human factor is also problematic in deploying classifiers for text-based analysis. While classifiers can be trained to detect patterns and identify potential solicitation, they cannot accurately replicate human communication's nuances. This means that there is a risk of misinterpreting communication, which can lead to false positives or, in extreme cases, the identification of innocent individuals as potential offenders. Human assessors must, therefore, be trained to interpret the results accurately and understand the limitations of the classifier.

Impact on the security of communications:

No additional points raised, see *Impact on Security of Communications* for ‘Classifiers for imagery’.

The Risk of repurposing:

No additional points raised, see *Risk of Repurposing* for 'Classifiers for imagery'.

4.4 ON-DEVICE FULL HASHING WITH MATCHING AT THE SERVER (E2EE COMMUNICATION)

Disclaimer: This section summarises the discussions amongst participants. The authors take responsibility for any inaccuracies in interpretation but are not responsible for the accuracy of the statements made during the workshop.

On-device full hashing with matching at the server is a detection method that involves creating a hash of every image and video on a device and then comparing these hashes to a database of known CSAM hashes. The image or video is flagged for further review if a match is found. This process occurs on the device, meaning the user's data is not sent to a central server for analysis. This method is designed to address concerns about privacy and security, but it is not without risks.

To ensure that the risks and impact of on-device scanning are accurately assessed, a clearer taxonomy of on-device full hashing with matching at server, also called 'client-side scanning' (CSS), is needed. This would involve defining different types of scanning, such as CSS with/without reporting to authorities, CSS with/without deletion of illegal content, CSS on the app, and CSS on the device. Each of these different scenarios would have different implications for privacy, security, and accuracy, and they must be analysed and evaluated separately to ensure that the appropriate measures are taken.

***Type of online CSA which can/cannot be detected**:*

On-device full hashing with matching at the server can be used to detect known CSAM. However, it cannot be used for the detection of unknown CSAM or heavily modified versions of known CSAM, nor can be it be applied in the detection of conversations indicating possible solicitation.

Method of compliance with detection order:

To comply with a detection order, platforms implementing on-device full hashing must connect to a database of known CSAM hashes. This database must be highly secure, and access should be limited to a few trusted individuals. Platforms must also develop mechanisms to create a hash list on the device without compromising user privacy and security. Once the hash list has been created, the platform can access an Application Programming Interface (API) to match the hashes with those in the database. This process must be highly secure, and the platform must ensure that the results are accurate and reliable. A human must review any images or videos flagged as CSAM to confirm they contain CSAM.

It must be noted that complying with a detection order by using on-device full hashing technology could create a significant burden on tech companies. These companies must store a database of known CSAM hashes, which must be updated regularly. They would also need to implement the technology on their platforms, which could be complex and resource intensive. This could lead to a situation where only large tech companies with significant resources can comply with the EU's proposal, while smaller companies lacking the technical, human, and financial resources cannot.

***Impact on confidentiality of communications**:*

There were different opinions amongst participants regarding the impact on confidentiality of communications. Some argued that this technology can be implemented without compromising user privacy. The system only creates a hash of each file, meaning that law enforcement only sees the hash, not the image or video itself. Additionally, the system only searches for known CSAM hashes, which limits the potential for false positives. It was argued that the technology could be deployed with surgical precision, and additional safeguards could be put in place to minimise the risk of compromising the privacy of all users. Platforms should also implement robust security measures to protect the hash list and prevent unauthorised access. Further, platforms must develop an appeals and review process that allows individuals to challenge the findings of the detection technology. This process must be transparent, accessible, and fair, allowing individuals to present evidence to support their claims.

Using on-device full hashing also raises concerns about the privacy and security of individuals. The technology requires access to a user's device to create a hash list, which could be seen as an invasion of privacy. To address this concern, some argue that platforms must ensure that users are fully informed about the use of detection technologies and that they can opt out of the process. In addition to these risks, the use of on-device full hashing also raises ethical concerns. The technology could treat children solely as objects of protection rather than fully formed subjects of rights. Platforms must implement detection technologies to recognise the full range of children's rights, including their right to privacy and freedom from surveillance.

Further, it was raised that platforms must ensure that the review process minimises the risk of further collateral damage to the victim, such as inadvertently notifying other family members without the victim's consent. This process must be designed to recognise children's rights as fully formed subjects rather than objects of protection.

Impact on the security of communications:

The impact on security was considered low by some. However, other participants stated that security is breached by the process and once a security breach has been created, it remains vulnerable to bad actors.

Another concern is the quality and accuracy of the CSAM databases used for comparison (reference is made to similar points being raised under section 4.1). To mitigate these risks, on-device full hashing with matching at the server must be implemented with robust security measures. These measures should include using strong cryptographic algorithms to reduce the risk of false positives due to hash collisions. Additionally, CSAM databases should be regularly updated and maintained to contain accurate and complete information.

The database security of hash values should also be a top priority. This should include encryption and access controls to protect against unauthorised access. Furthermore, auditing and monitoring tools can help detect and prevent cyberattacks. It is also crucial to use on-device full hashing with matching at the server in a transparent and accountable manner. This should include clear guidelines on how the technology is used and how the data collected is stored, shared, and protected. The public should also have access to information on the accuracy and completeness of the CSAM databases and the measures taken to ensure their security.

Another point raised was that the CSAM database could be a target for cybercriminals, who may attempt to gain access to it to identify the flagged file users. This could result in the exposure of sensitive personal information, such as the identity of individuals who possess CSAM.

Lastly, the scanning software would need to be integrated into operating systems, which could create vulnerabilities that could be exploited by hackers and malicious actors. Additionally, constantly scanning users' devices could lead to increased battery drain and slower performance, negatively impacting the user experience and raising concerns about data usage.

Risk of Repurposing:

The risk of repurposing this technology was considered high. This is because manipulating the system on the device is a considerable risk, and a constant update of high security would be needed to guard against attacks.

It is also essential to consider the potential impact of on-device scanning on freedom of expression and privacy rights. While detecting CSAM is a noble goal, any measures taken must not infringe on individuals' privacy and free expression rights. For example, there is a risk that the scanning technology could be repurposed for surveillance purposes, or that legal content could be censored or restricted based on overly broad definitions of what constitutes CSAM (reference is made to similar points being raised under section 4.1).

Ensuring that the technology is subject to robust oversight and accountability measures is vital. This could include regular audits of the system to ensure that it is being used only for its intended purpose and that user privacy is respected. Additionally, there should be clear procedures for handling false positives and ensuring innocent users are not falsely accused of illegal activity.

4.5 ON-DEVICE PARTIAL HASHING WITH REMAINING HASHING AND MATCHING AT THE SERVER (E2EE COMMUNICATION)

Disclaimer: This section summarises the discussions amongst participants. The authors take responsibility for any inaccuracies in interpretation but are not responsible for the accuracy of the statements made during the workshop.

Type of online CSA which can/cannot be detected:

On-device partial hashing with remaining hashing and matching at the server within E2EE environments can only detect known CSAM.

Method of compliance with detection order:

To comply with a detection order as outlined in the proposed Regulation, a company would need to build a system on the device that is highly secured and has a built-in server system, before implementing three main actions: 1) Establish a connection to a database of known CSAM; 2) install on device a system to create partial hashes of images and videos; 3) using an Application Programming Interface (API), establish a connection to a server of known CSAM to perform matching. In the case of a match, a report would need to be generated and referred to the competent authorities. Access to

metadata would also be required, as would a smaller number of human moderators with the training and mandate to determine the legal status of the material in question. An appeals and review process would also be essential in the case of false positive or concerns about any part of the process.

Concerns related to this method include practical difficulties for companies in determining the issuance of orders, if the user is in another jurisdiction, oversight to safeguard individual privacy rights, and the risk of misuse of the system by malicious actors.

Impact on confidentiality of communications:

The impact was considered lower compared to classifiers used in open communications, but still relatively high. Concerns include access to private and sensitive data during the detection process, and the requirement to deploy detection technology on locally stored (on-device) images in addition to images stored on a cloud device or messaging service. An additional concern relates to malicious actors' potential misuse, which must be carefully mitigated through high-security standards.

***Impact on the security of communications**:*

The impact on security was considered low by some. However, other participants stated that security is by default breached by the process because once communication is open, it remains open and vulnerable to hacking and cyberattacks.

***Risk of repurposing**:*

The risk of repurposing the use of a technology or data for purposes other than those for which it was intended was considered high. The proposed Regulation is intended to be technology neutral, which without greater clarity could require a company to develop and/or deploy a secure technology to carry out on-device partial hashing with remaining hashing and matching at the server.

Compliance with detection orders has the potential to be complex and challenging and may result in the compromise of personal data that would need to be rigorously monitored.

4.6 SECURE ENCLAVES IN THE ESP SERVER (E2EE COMMUNICATION)

Disclaimer: This section summarises the discussions amongst participants. The authors take responsibility for any inaccuracies in interpretation but are not responsible for the accuracy of the statements made during the workshop.

Secure enclaves are designed to protect against attacks that exploit vulnerabilities in the operating system or application code. Secure enclaves have a significant impact on the confidentiality of communications within an ESP server. By isolating sensitive data and computations within enclaves, secure enclaves ensure that even if the system is compromised, the data and computations remain secure. This protection is critical in preventing the interception of communications and the theft of sensitive data.

NOTE: Due to time limitations, there was no extensive discussion around the use of secure enclaves to carry out detection in E2EE environments. However, several discussion groups did provide

comments in note form about security, confidentiality, and compliance with a detection order. The authors have added some explanatory text for the reader.

***Type of online CSA which can/cannot be detected**:*

The use of secure enclaves would be possibly only in the case of detecting known CSAM.

Method of compliance with detection order:

Like the other methods for detection in E2EE, compliance with a detection order using a secure enclave would require a company to build a system on the device that is highly secured and has a built-in server system and a mechanism to hash on device. They would also need access to a hash database, an API access point, and to build a system on the device that is highly secured and has a built-in server system. Access to metadata would also be required, as would a smaller number of human moderators with the training and mandate to determine the legal status of the material in question. An appeals and review process would also be essential in the case of false positives or concerns about any part of the process.

***Impact on Confidentiality of Communications**:*

The impact on Confidentiality of Communications of detection in E2EE using a secure enclave was considered medium to high. This is because secure enclaves protect the confidentiality of communications between authorised parties, thereby ensuring that sensitive information remains secure and only accessible to those authorised parties. However, while security of communications may be considered very strong using secure enclaves, confidentiality depends on the ownership and possession of the key for the encrypted communications.

***Impact on Security of Communications**:*

Secure enclaves also have a significant impact on the security of communications within an ESP server. By creating isolated environments within the system, secure enclaves can provide a high level of protection against attacks that attempt to compromise the system or steal sensitive data. As such, secure enclaves have strong potential for use to detect CSAM, for example when distributed using P2P systems like email and messenger services. By using secure enclaves to analyse content, ESP servers can also be used to identify conversation patterns and behaviours associated with the solicitation of children for sexual purposes. This analysis can be performed within the enclave, ensuring that the data and computations remain secure and are not accessible to unauthorised parties.

In addition to analysing content, secure enclaves can identify and block CSAM before they are delivered to users, in the same way that they are already used to analyse attachments and URLs for malware and phishing attacks. This analysis can be performed within the enclave, ensuring that the data and computations remain secure and are not accessible to attackers. Secure enclaves can also provide a secure environment for running antivirus and anti-malware software.

For these reasons, and overall because secure enclaves are a powerful technology that can provide a high level of protection for data and computations within a system. the impact on security of communications of detection in E2EE using a secure enclave was considered low.

***Risk of repurposing**:*

The risk of repurposing when detecting in E2EE using a secure enclave was considered high due to the potential for actions other than those originally intended, such as to detect CSAM. However, it is noted that risk does not equate to ease or practicality; strong safeguards are needed to guard against the risk that does exist.

5. RECOMMENDATIONS

The workshop participants agreed that the proposed Regulation raises complex questions about privacy and protection, and it is crucial to strike a balance between these two aspects. Participants agreed that E2EE technologies must be preserved while ensuring the proposed Regulation's objectives are achieved. The recommendations made during the workshop can help in safely deploying detection technologies to protect children from abuse while protecting fundamental rights and ensuring accountability and transparency in the process.

5.1 RECOMMENDATIONS FOR LESS INTRUSIVE MEASURE FOR REMOVING CSAM

- **Define different objectives of proposed Regulation**: Participants agreed on the importance of distinguishing between removal of CSAM from the internet and processes for identifying victims including by investigating online CSA offenders – actions outlined in the proposal would require platforms to ensure the former and facilitate the latter via reporting and referral processes.

- **Detect CSAM without reporting the ID of the sender:** pursuing the removal of CSAM can be done without revealing the identity of the person who shared the imagery; some participants recommended a suppression model (detecting and removing without reporting) as a starting point until more effective privacy-preserving technology is ready (staggered approach).

- **Detect CSAM without reporting in E2EE platforms:** Some participants suggested that client-side scanning of known CSAM within E2EE and removing such content without informing law enforcement agencies could be a possible way to preserve privacy. If the imagery is flagged on a device because of hashing, the identity of the user would not necessarily have to be indicated. However, it must be noted that this does not assist in advancing further investigations into active cases.

5.2 RECOMMENDATIONS FOR E2EE AND DETECTION TECHNOLOGIES

- **Acknowledge differences between E2EE platforms:** Participants highlighted that not all platforms using E2EE should receive the same risk rating. The risk associated with E2EE technology depends on the service's context: for example, WhatsApp, which is E2EE, might have a lower risk than Messenger on Facebook, which allows contact with potentially anyone.

- **Recognize that E2EE not a goal in and of itself:** Participants highlighted that the strong focus on safeguarding the E2EE communication might be misleading as protecting E2EE technology should not be a goal in and of itself. The main objective behind it is to ensure the confidentiality of communication. Confidentiality not only serves the privacy of

communication, but it also enables the exercise of other fundamental rights such as freedom of speech.

- **Preserve the high societal value of E2EE:** Participants discussed the clear societal value of E2EE, largely supporting the notion that creating exceptional access to E2EE platforms would undermine this value and represent a considerable security risk for E2EE communications overall. As such, alternatives to creating a 'back door' should be prioritised.

- **Exercise caution over client-side scanning to avoid creating the conditions for a broader surveillance infrastructure:** In addition, some participants supported the notion that client-side scanning on device, while it has benefits for the detection of CSAM and the confidentiality of user data, also presented risks as this creates a technological infrastructure that could be used for broader monitoring of data and communications on consumer devices, which would negatively impact upon the fundamental rights of device users by creating a platform or app-agnostic method to monitor data. As such, caution and the creation of robust safeguards and oversight was recommended for this type of technology.

5.3 RECOMMENDATIONS FOR ADDITIONAL LEGAL SAFEGUARDS FOR DETECTION ORDERS AND DETECTION TECHNOLOGIES

- **Enable a staggered approach to technology deployment by certain companies:** There was interest in and support for a staggered approach to mandatory detection, particularly in the case of smaller online service providers, those with a low-risk assessment, and those that have never previously deployed detection technologies. The approach of 'asymmetric obligations' is set forth in the Digital Services Act[11] and should similarly apply to the proposed Regulation. This should be clearly articulated in the proposed Regulation's text addressing risk assessment, mitigation, and detection orders.

- **Provide more clarity on risk assessment and mitigation processes: It is essential to provide clarity and guidance on how the risk assessment process** will be carried out. To assess risk, companies would need to consider using hash-based detection technologies as part of a suite of assessment measures such as user reporting, content moderation, and random checks. The proposed Regulation should therefore provide minimum guidelines for risk assessment, including the legal basis where necessary.

- **Introduce stronger legal prohibitions to prevent repurposing:** Stronger legal prohibitions should be added to the proposed Regulation to ensure that detection technologies are never repurposed for other forms of detection in private communications beyond online CSA. The current provision in Art 10 (4), which addresses the risk of repurposing, puts this obligation on the provider. However, there should be a legal commitment from the EU that the detection regime under the proposed Regulation will not be expanded to other illegal or harmful content or used for commercial purposes.

[11] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act).

- **Introduce guidance on content moderation and human review:** Participants also recommended that the proposed regulation should include guidance on outsourcing content moderation without violating data protection and privacy laws and being criminally liable for the content. Considering the potential for a huge increase in communications requiring human review due to the proposed Regulation, it is crucial to avoid that such human review is outsourced without authorisation or oversight outside the EU, where working conditions for content moderators might be particularly poorly regulated and lead to negative health and wellbeing outcomes.

- **Provide clarity on the framework for testing detection technologies:** Participants noted that the detection technologies would have to be tested and evaluated before their accreditation and deployment. It was assumed that EUROPOL as the sister agency of the EU Centre would be mandated to conduct and oversee testing. This responsibility for testing should be clarified in the proposed Regulation to avoid the creation of multiple frameworks.

- **Ensure transparency around data sets and avoiding classifier bias:** Classifier bias is an important consideration, particularly if classifiers are trained on datasets that lack representation, for example from different racial groups. There is currently an insufficient representation of diverse genders, races, and age groups in the training data set. Therefore, the false negative detection rate, i.e., the imagery falsely not flagged as CSAM despite being CSAM, will disproportionately affect children underrepresented in the training data set. The classifier might hence disproportionately detect more children who fit into the categories mainly prevalent in the training data set, leaving children underrepresented in the data set without protection. This is a strong argument for addressing obstacles to accessing data between and across sectors for the purposes of classifier training. The EU Centre can play a crucial role in addressing this challenge.

- **Specify the role of hotlines in providing data sets:** Hotlines are currently not specifically mentioned even though they will play a crucial role in providing the data sets to the EU Centre.

- **Introduce mandatory human rights and child rights impact assessments during the testing of detection technologies:** The participants recommended that human rights impact assessments, with a particular focus on children's rights, should be considered as part of the technology assessment process. Children's and survivors' voices should also be included in the life cycle of the technology assessment and deployment process.

- **Enable use of multi-version of solicitation classifiers:** Participants highlighted different types of solicitation, and each platform has a different approach to solicitation detection. Therefore, it is crucial to have multi-version solicitation classifiers to detect different types of solicitation. For example, classifiers may be designed to identify solicitation aimed at children of different ages or involving a request to meet in person. This can help platforms tailor their approach to solicitation detection and improve detection accuracy.

- **Ensure clear appeal and review process:** An appeal and review process are necessary to ensure the ethical use of classifiers for text-based analysis. This process should involve a human review of any content flagged by the classifiers. Moreover, a legal basis for such review

should be established, and the appeal process should involve a review by independent experts. However, such a process is human-intensive and requires the employment of more humans to carry out the review.

- **Provide access to accredited technologies and independent auditing:** To ensure the accuracy of classifiers for text-based analysis, it is essential to have access to accredited technologies. These technologies should be developed according to industry standards and regularly updated to keep up with the latest trends in the solicitation of children for sexual purposes. This should also include independent auditing of such accreditation, which must be reflected as a legal obligation in the proposed Regulation to increase transparency.

- **Enhance accuracy of language analysis through use of metadata and context of conversations:** To improve the accuracy of classifiers for text-based analysis, the metadata and context of the conversation should be considered. This can help identify patterns and context-specific language that may indicate solicitation. For example, the conversation's time and location may indicate solicitation activity, and the use of specific words or phrases may also be suggestive of such activity. Therefore, metadata and context should be incorporated into classifiers to improve accuracy. However, this means that even more metadata and communications data will be collected, which impacts the right to privacy.

- **Replicate good practice for classification:** To ensure safeguards around the verification of solicitation, a triple-vetting, or 'three eyes' model as used by some hotlines and law enforcement in particular for CSAM could be adopted. This can help reduce false positives and increase the accuracy of detection. However, it is noted that this process is time-consuming and resource-intensive and may not be feasible for all platforms.

- **Be guided by existing ontologies and best practices:** To ensure the ethical and legal use of classifiers for text-based analysis, the method of compliance should be guided by existing ontologies and best practices within the industry. This can help ensure the accuracy and effectiveness of the classifiers and prevent their misuse. Moreover, adherence to existing ontologies and best practices can help establish trust between platforms and users, which is essential for effectively detecting solicitation.

5.4 RECOMMENDATIONS FOR TRANSPARENCY AND OVERSIGHT OF TECHNOLOGY COMMITTEE

- **Ensure a multi-disciplinary Technology Committee (TC):** Participants recommended that the TC be a multi-stakeholder body as Article 66 (1) currently does not set any specific areas of who the experts on the committee should be. The TC should include experts from different fields, such as linguistics, psychology, computer science, privacy, child protection and others. Survivor representation should also be considered for the TC.

- **Carve out a more active role for the Technology Committee**: Participants recommended that the TC should be able to act based on their initiative, not only when requested by the EU

Centre as currently stated in Art 66 (6) (c). Such action could include the development of Guidelines and Codes of Conduct, issuing opinions and putting specific issues in need of clarification to the EU Centre. The TC should develop documentation standards for the detection technologies approved by the EU Centre. This could include guidelines on using classifiers to ensure their ethical and legal use. This is especially important given the potential for false positives and their consequences for individuals wrongly accused as perpetrators.

- **Guarantee transparency of proceedings:** The proceedings of the TC should be on open record to increase accountability and trust in the body.

- **Develop minimum viable accuracy standards for detection technologies:** The Technology Committee is encouraged to discuss, publish and be transparent about the minimum viable accuracy standards required for any detection technology. No detection technology should be made available without robust independent compliance verification with this standard. This standard should be applicable for both false positives and negatives.

5.5 RECOMMENDATIONS FOR CLEAR ROLE FOR PARTICIPATION OF CHILDREN AND SURVIVORS

- **Give a clear role to voices of survivors in the proposed Regulation:** participants raised that survivors are currently not represented in the proposed regulation; voices of survivors are also diverse: some survivors are in agreement with client-side scanning and monitoring of communications, others are against such measures and stress the importance of high levels of privacy; this should also include survivors outside of the EU whose imagery is circulating within the EU.

- **Give a clear role to voices of children in the proposed Regulation:** Children need to play an active role both in the development and in the implementation of the proposed Regulation; protection of children and privacy are interrelated, and children should not simply be considered objects of protection.

- **Actively mitigate risk of harm to children through interference with their privacy:** Interference of children's privacy might lead to an increasing risk of violence; potential harm might be caused by the investigation of online CSA where such material was produced and shared voluntarily.

- **Create pathways for easily accessible and anonymous reporting:** If children report certain instances anonymously, participants agreed this anonymity should not be breached; children who report such cases need to be able to remain in control of the process; for example, there should be pathways for children to anonymously add their imagery to existing hash list at the EU Centre.

6. CONCLUSION AND WAY FORWARD

The technicalities of the proposed Regulation and their impact on fundamental rights of adult and child users were the topic of discussion at the second expert workshop held at VU-Amsterdam. The key topics discussed were the use of hashing technologies, classifiers for imagery and text, and other detection technologies for both open and E2EE environments. The risk of repurposing these technologies is a crucial topic which requires strong legal safeguards. Further, careful consideration of how specific data sets are used is necessary and how data bias can be avoided to provide the same level of protection to all children. The protection of E2EE is not a purpose in and of itself: the integrity of communications, through E2EE and other technologies, is vital for protecting not only the right to privacy, but also other rights such as freedom of expression. Lastly, the proposed Regulation needs to find a platform for hearing the voices of survivors and children as a cornerstone of the development and the implementation of the proposed Regulation.

The proposed Regulation is a relevant step towards addressing the issue of online CSA. Children are particularly vulnerable to the devastating impact of online sexual abuse and exploitation, and the proposed Regulation includes several provisions aimed at ensuring their protection and promoting their rights. Accordingly, an additional workshop is planned to discuss the impact of online CSA on children and the proposed Regulation's compliance with children's fundamental rights. The following is a non-exhaustive list of topics under consideration for discussion at a third Expert Workshop:

- **Age Assurance/Verification:** One of the critical measures included in the proposed Regulation is age assurance/verification. Ensuring that children are not exposed to online CSA is an essential step in protecting them from harm. At the same time, the right to privacy and personal data protection needs to be considered for both children and adult users to avoid violations of these rights through age assurance/verification measures.

- **Consensual Sexual Exploration:** Another critical issue that needs to be addressed in the proposal and contextualised in the EU legal landscape in particular Directive 2011/93 is consensual sexual exploration. Children should be able to explore their sexuality safely and without harm. The proposed Regulation requires platforms to take measures to prevent online CSA which might have negative impact on children who consensually explore their sexuality online, such as having their imagery/conversations flagged to law enforcement for investigation.

- **Child-Friendly User Reporting:** Child-friendly user reporting is also a crucial element that needs addressing in the proposed Regulation. Children who encounter harmful content online should be able to report it quickly and easily. Platforms should provide child-friendly reporting mechanisms to encourage children to come forward and seek help.

- **Child Rights Impact Assessments:** Workshop Participants strongly urged provisions for independent child rights impact assessments. This will ensure that the proposed Regulation does not inadvertently harm children's rights while attempting to protect them. The lack of sufficient data for Europe on children's voices must also be addressed in developing and implementing laws and policies.

- **Victims' Rights Directive:** The regulation interacts with the victim rights directive. The proposal includes provisions to ensure that victims of CSAM are adequately supported and their rights are respected. The proposal recognises the importance of the EU Centre in helping victims of CSAM, particularly young offenders.

- **Child Protection Ecosystem:** The broader child protection ecosystem is crucial in addressing online CSA. The proposed Regulation recognises the importance of addressing the incentives for certain groups to engage in online CSA for economic purposes, including organised crime purposes, the agency of children in exploitation, sexual extortion, and the involvement of families facilitating crimes.

- **Cross-Border Aspects:** The proposal should better reflect the cross-border aspects of online CSA, which may impact different markets differently, depending on the risks associated with deployment in different EU countries.

- **Design Discrimination:** Another critical issue is design discrimination. Children in different legislations may enjoy different levels of protection on platforms. The regulation aims to ensure that all children are protected regardless of location, recognising that children's rights should be universal and not subject to geographical boundaries.

In conclusion, the proposed regulation is a significant step in promoting children's rights and protecting them from harm. However, it is essential to recognise that external safeguards cannot only be included in the proposed Regulation but must be considered for the broader ecosystem. These include education campaigns, increased collaboration between law enforcement agencies and technology companies, and international cooperation to address the global nature of online CSA. The proposed Regulation includes several provisions to ensure their protection, and these measures must be implemented effectively to achieve the desired outcomes.

We would like to thank all participants for their constructive and collegial engagement during this Expert Workshop. All participants of the workshop were in attendance because of their unwavering believe in the importance of finding proportionate solutions to online CSA that respect and protect the fundamental rights of all users - children and adults alike. We hope that future workshops will continue to serve as a platform for knowledge sharing, critical reflection and learning and help to develop evidence-based, informed EU laws and policies.