Council of the European Union
General Secretariat

Brussels, 26 February 2024

Interinstitutional files:
2022/0155 (COD)

WK 3036/2024 INIT

LIMITE

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

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

WORKING DOCUMENT

From:	Presidency
To:	Law Enforcement Working Party (Police)
Subject:	Proposal for a Regulation of the European Parliament and of the Council laying down rules to prevent and combat child sexual abuse - possible criteria and classification methodologies for the risk categorisation of services

The Presidency provides delegations in the Annex with a working document outlining possible criteria and classification methodologies for the risk categorisation of services to facilitate discussions at the meeting of the Law Enforcement Working Party (Police) on 1 March 2024.

ANNEX

Proposal for a Regulation of the European Parliament and of the Council laying down rules to prevent and combat child sexual abuse – possible criteria and classification methodologies for the risk categorisation of services

As expressed in the discussion paper 6850/24, the Presidency suggests developing a methodology for determining the risk of specific services, taking stock of the risk assessment and risk mitigation measures outlined in Articles 3 to 5b of the latest compromise text of the proposed regulation.

The services would be categorised into 4 different levels of risk (negligible, low, medium and high) which would have consequences for the providers concerning the obligatory or voluntary application of mitigation measures, the possibility of receiving detection orders, and the frequency of the regular re-assessment of the risks.

To allow an objective risk categorisation, the operative text should encompass concrete and definite procedures based on a set of criteria that must reflect, as precisely as possible, the potential harmfulness of a service with regard to child sexual abuse. These criteria should be assessable and concretely observable and could, for instance, be related to the type of service, the core architecture of the service, the provider's policies and safety by design functionalities, and user tendencies.

While the criteria to be applied and the principles of the methodology for the risk categorisation have to be laid down in the legislation, technical details could be left for implementing acts. It is also relevant to keep agile with regard to adapting the risk categorisation to future technological developments, and to define the roles that the Commission and the EU Centre could play in supporting the risk categorisation process.

In order to steer forthcoming discussions on the risk categorisation and details of scoring systems, the Presidency is proposing in this working document a non-exhaustive list of criteria that could be used as inspiration and illustrate the direction that such a procedure could take. It should be stressed, therefore, that this living document is intended to provide examples reflecting the state of progress of Presidency reflections.

The Presidency's idea is to identify, as extensively as possible, the potential criteria that could serve as risk denominators and to consider the classification methods that could be envisaged.

Delegations are therefore invited to consider these proposals, to comment on their relevance and to propose possible other criteria and methodologies.

I. Possible risk categorisation criteria

1) Based on the category of services

A first approach could be to distinguish the potential risk if the service provides one or more of the following service types.

- Social media platform (services that connect users and enable them to build communities around common interests or connections)
- Electronic messaging service (services that are typically centred around allowing users to send messages that can only be viewed or read by a specific recipient or group of people)
- Online gaming service (services that allow users to interact within partially or fully simulated virtual environments)
- Adult service (services that are primarily used for the dissemination of user-generated adult content)
- Discussion forum or chat room service (services that allow users to send or post messages that can be read by the public or an open group of people)
- Marketplace or listing service (services that allow users to buy and sell their goods or services)
- File-storage and file-sharing service (services whose primary functionalities involve enabling users to store digital content and share access to that content through links)

2) Based on the core architecture of the service

This section assesses the level of interaction which is possible between users on a service.

a) Access for children to a part or the entirety of the service.

b) User identification functionalities

- to display information through a user profile that can be viewed by others (e.g. images, usernames, age).
- to share content anonymously (e.g. anonymous profiles or access without an account).
- To have a multi-factor authentication and user sign-up information gathering (phone number, email address, or other identifiers).

c) Connection methods for users to reach other users[1]

- Possibility to form closed groups and/or send group messages.

[1] 'User connections' (UK Safety Act): A user-to-user service functionality that allows users to follow or subscribe to other users. Users must sometimes be connected to view all or some of the content that each user shares.

d) Possibilities on user communication

- Communication via livestream.
- Communication via direct messaging/ ephemeral direct messaging.
- Communication via encrypted messaging and functionalities for "Opt-in/ opt-out"[2]
- Commenting on content (either open or closed channels).
- Posting/ sharing images or videos (either open or closed channels).
- Posting/ sharing location information.
- Reposting and forwarding content (either open or closed channels).
- Searching for user-generated content.
- Sharing of content via hyperlinks and plain-text URLs.

e) Possibility for users to post goods/services for sale[3]

3) Based on policies and safety by design functionalities in place

This section assesses the level of measures taken by the providers to protect child users.

a) Effectiveness of CSA Risk Policies

b) Strength of Prohibitions and Restrictions

c) Functionalities for Age Verification [4]

- Privacy Protection: The age verification system protects user privacy, ensuring data is not disclosed or processed for any purpose other than age verification.
- Minimal Data Collection: Only minimum of data is collected, adhering to a minimal data collection approach for age verification.
- Data Retention: Personal data related to age verification is not retained after the verification process is completed.

[2] Making design choices such as whether the use of E2EE is opt-in by default, rather than opt-out which would require people to choose E2EE should they wish to use it, therefore allowing certain detection technologies to work for communication between users that have not opted in to E2EE. Links to encrypted services are often shared on unencrypted online spaces to facilitate the exchange of CSAM.

[3] 'Posting goods and services' (UK Safety Act): a user-to-user service functionality allowing users to post content dedicated to offering good and services for sale. This does not include paid-for-advertisement but may serve the function of allowing users to promote goods and services. Potential perpetrators may try to promote illegal goods or services by posting them for sale using this functionality. Often illegal items such as drugs and firearms are posted for sale using code names. In certain contexts, the ability to post goods or services for sale, such as through user-generated advertisements, also enables potential perpetrators to advertise and broadcast the sexual services of adults in exploitative environments. The risk of harm can be increased if services also allow users to make online payments directly.

[4] A provider is only entitled to conclude that it is not possible for children to access a service, or a part of it, if age verification or age estimation is used on the service with the result that children are not normally able to access the service or that part of it.

- Proportionality: The age verification system is proportionate to the risks associated with the product or service, ensuring a balanced approach.
- Remedies and Redress: Adequate remedies and redress mechanisms are provided for users whose age is wrongly identified.
- Selective Disclosure: Users have the option for selective disclosure of attributes during the age verification process.
- Zero-Knowledge Protocol: The system requires the user to provide a 'token' to confirm that the minimum age requirement is met.[5]
- Anonymous Accounts: Users have the option to use anonymous accounts for age verification, enhancing privacy protection.
- Non-Identification Requirement: The age verification system does not require the identification of each user of a service.
- No Biometric Data Processing: The system does not require the processing of biometric data during the age verification process.

d) Functionalities for Parental Control Scoring System

e) Functionalities for Notifying/flagging Online Child Sexual Abuse

f) Efficiency in Handling Notified/flagged Potential Child Sexual Abuse

g) Statistical Information Completeness and Relevance

h) Measures for Promoting Users' Media Digital Literacy and Safe Usage Scoring System [6]

i) Alignment of Business Model, Governance and Systems with CSA risk mitigation

j) Functionalities enabling users to Share Potentially Harmful Content

k) Functionalities Assessment of Potential Dissemination Risks

4) Based on user tendencies and statistics

This section assesses the user tendencies and trends based on a statistical analysis of users.

a) Assessing User Patterns

b) Service's Popularity Among Different Age Groups

c) Analysis of Solicitation Risks Based on User Mapping

---

[5] Such a token is produced once a person's age is verified or estimated by an age assurance provider and allows the online service to confirm that the age requirement is met without viewing or collecting users' personal information. The token can be stored in a device's digital wallet or browser to be reused for a period of time to access services requiring the same level of assurance.

[6] Duties about children's access assessments:
- A provider must carry out the first children's access assessment within one year.
- A provider must carry out a children's access assessment of the service:
* before making any significant change to any aspect of the service's design or operation to which such an assessment is relevant,
* in response to evidence about reduced effectiveness of age verification or age estimation that is used on the service,
*in response to evidence about a significant increase in the number of children using the service.

d) Related to Account:

- Use of Anonymous Account
- Consecutive and Repetitive De- and Re-Activation of Accounts
- Fake or Imposter Accounts
- Identity Verification Tools for Opening Accounts
- Pseudonymity
- Temporary Accounts
- Frequent Changing Accounts of Profile Details
- Unmatching of Defriending Victims on Social Media Accounts
- Switching Between Private and Public Platforms
- Moving Public Conversation to Private Channels
- Obfuscation of I.P Addresses
- Use of Unsecure Public WIFI Hotspots
- Creation of Private Group or Chatboxes
- “Cyber Flashing” tendency (Unsolicited Intimate Messages)

5) Related to Company Policy on User Safety

This section assesses the measures implemented by the service to ensure the safety of its users.

a) Usage of Premoderation functionalities

b) Usage of Delisting Content System

c) Usage of Image Masking

II. Possible scoring methodologies

The risk categorisation system would be based on a set of parameters to which different scoring methodologies could apply, such as binary questions, hierarchical criteria (Absent / Basic / Effective / Comprehensive), or sampling as currently proposed in Article 47a of the latest compromise text. The procedure could, if relevant, also integrate a combination of these solutions:

1) Binary methodology

Scoring based on simple yes/no questions related to the core architecture of the service. A yes/no response could be given a +/- score respectively (meaning more/less risk) resulting in a final score.

For example: Does the service have a livestream system? A "yes" will represent a positive value whereas a "no" will represent a negative value. A higher total score implies a higher risk. The potential range of possible scores will then be divided into 4 categories accordingly.

2) Multi-class scoring with 4 hierarchical criteria methodology

Scoring based on the extent to which policies and functionalities are in place to address the risk of child sexual abuse material being disseminated or grooming practices taking place on the service. Scoring could be based on 4 levels: Absent / Basic / Effective / Comprehensive. Each level would represent a score from 4 (high risk) to 1 (low risk) resulting in a final score.

For example:

Functionalities for notifying/ flagging Online Child Sexual Abuse

- Absent/very limited
  - The platform lacks or has very limited functionalities enabling users to flag and report online child sexual abuse in the sense that they are ineffective.
- Basic
  - The platform provides basic flagging tools, but accessibility and age-appropriateness need improvement. Need to enhance the accessibility of reporting tools to ensure users can easily locate and utilise them. Might improve the interface to make reporting tools more age-appropriate, especially for younger users.
- Effective
  - The reporting tools are effective, offering users a straightforward and age-appropriate means to flag and report online child sexual abuse. Ensure reporting tools are easily accessible within the platform, promoting quick and efficient reporting. Maintain an age-appropriate interface for reporting tools, catering to users of all age groups. Provide ongoing educational resources to keep users informed about recognising and reporting online child sexual abuse.
- Comprehensive
  - The platform excels by providing comprehensive tools for notification of online child sexual abuse, ensuring a swift and effective response to flagged content.
  - Collaborate with external organisations and law enforcement agencies to enhance the efficiency of the reporting and response process.
  - Regularly update reporting tools based on user feedback and technological advancements.

3) Sampling methodology

For certain evaluation criteria and parameters, it may also be possible to implement a compartmentalization system based on the sampling and analysis of specific data. This is a relatively technical method which, if used, requires specification of the types of data to be sampled, the collection procedures, the compartmentalization mechanisms, and so on. It is the same type of method that is defined in article 47a and article 7(2) of the latest compromise text proposed under the former Spanish Presidency (12611/23).

For example:

This type of methodology could be used to analyze CSAM data itself in order to assess the risk, or metadata such as data relating to user accounts to assess the frequency of occurrence of certain risk-relevant points. This can be useful for calculating, for example, the extent of anonymous account use, the use of fake accounts, the frequency of account changes, the use of VPNs, etc. It would therefore be necessary to collect the data from providers, process it, and conclude trends based on these samples.

Use of Anonymous Account:

- *frequent use of anonymous accounts*
  - *Over 25% of accounts lack identifiable information.*
- *Moderate instances of anonymous accounts.*
  - *From* 26 to 60% *of anonymous accounts.*
- *Minimal or no use of anonymous accounts*
  - *Majority of account have identifiable information* (from 61% to 100%).