REPHRAIN

Protecting citizens online

REPHRAIN:

Scoping the Evaluation of CSAM Prevention and Detection Tools in the Context of End-to-end encryption Environments

Claudia Peersman, Emiliano De Cristofaro, Corinne May-Chahal, Ryan McConville, José Tomas Llanos and Awais Rashid.

Version 1.1 - July 2022

UK Research and Innovation · University of Bristol · UCL · The University of Edinburgh · King's College London · University of Bath

Scoping the Evaluation of CSAM Prevention and Detection Tools in the Context of End-to-end-encryption Environments

Version 1.1

July 8, 2022

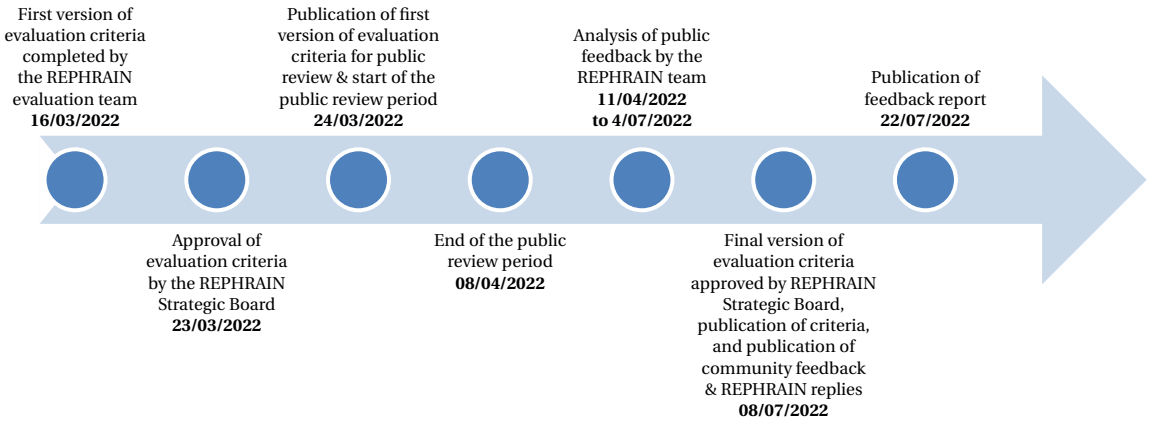
1 Summary

This document describes the scoping stage of REPHRAIN's independent evaluation of Proof-of-Concept tools for preventing and detecting child sexual abuse media (CSAM) within end-to-end-encryption (E2EE) environments that are currently being developed within five projects funded by the Safety Tech Challenge Fund[1] (the PoC tools). Given the tensions that arise between protecting vulnerable users, such as children, and protecting user privacy at large, key steps in REPHRAIN's evaluation process are (1) to seek input from the community, and (2) to publish all results, ensuring that academic rigour and objectivity remain at the core of our work, and to inform future directions in this area.

The evaluation criteria that have been developed are aimed to be a resource for the community and by the community. Hence, we invited feedback from members of the cyber security & privacy community along with stakeholders from academia, industry, law enforcement, and NGOs working in the field of online child protection. The community feedback phase ran for approximately 2 weeks (from 24 March 2022 until 8 April 2022).

The formal feedback request was published on the REPHRAIN website and circulated to the REPHRAIN contact list to ensure maximum exposure. Community feedback could be submitted via an online form where all comments were logged or could be sent via email, either as a free form text or an annotated PDF document.

The community feedback was reported to the REPHRAIN Evaluation Team for full consideration and discussion, and the REPHRAIN Strategic Board was advised accordingly. The final version of the evaluation criteria are published in this document on the REPHRAIN website, along with a summary of the key changes made (see Section 1.1. A full report including all feedback provided by the community will be made available shortly. This will detail how requests for changes were addressed, any changes made, and the rationale if a change was not made. The timeline for developing the evaluation criteria was adjusted as follows:



[1] https://www.safetytechnetwork.org.uk/innovation-challenges/safety-tech-challenge-fund/

1.1 Key Changes

The REPHRAIN evaluation team wishes to express its appreciation for the feedback we received from the community. We provide a summary of the key changes that were made in this section.

Compliance with human rights. The REPHRAIN evaluation team has been expanded to include a human rights expert, Dr. José Tomas Llanos (see Section 2.3), who has kindly agreed to contribute to a human rights impact assessment as part of the evaluation process. This extended scope (see also Section 1.2) led to the inclusion and enhancement of the first two evaluation criteria, which focus specifically on the impact of the proposed PoC tools on human rights and the extent to which their design is human-centred. The right to privacy was detached from the Security criterion and is now part of the Human Rights Impact criterion (see Section 3).

The definition of end-to-end encryption. A key element of the feedback we received referred to the debate about whether E2EE should preclude any access to information about the content of communications. As stated in the previous document, the REPHRAIN evaluation team does not wish to take a position on this definition. Hence, we removed all content that potentially suggested otherwise and clarified the scope of the evaluation in Section 1.2 accordingly.

Stronger focus on the effectiveness of the PoC tools on safeguarding children. A number of smaller changes have been made to help convey our general focus on how effective the PoC tools are in preventing and/or detecting CSAM, regardless of whether the tools employ AI-supported or hash-based technologies. This also allows us to evaluate how any additional functionalities of the PoC tools (e.g. pornography detection) interfere with their CSAM detection or prevention capabilities.

Auditability. This aspect was added to the fifth criterion (now called "Explainability, Transparency, Auditability and Provenance") — rather than being mentioned in the context of CSAM detection systems depending on matching a database of known content — because we agree that auditability is essential for any CSAM detection or prevention system.

1.2 Revised Objectives and Scope

The REPHRAIN evaluation team aims to provide a **technical assessment** that also takes into consideration the potential **implications for human rights** of each of the five proposed Proof-of-Concept tools based on the finalised version of the evaluation criteria presented in this document, while also contributing to the community debate regarding where potential challenges may lie with regards to privacy in an end-to-end-encryption framework, in the highly challenging context of online child protection.

The team is aware of the on-going debate about the definition of end-to-end encryption[2]. We do not wish to take a position on this definition within the framework of this study. Hence, we will be referring to the task at hand as evaluating technologies being applied within E2EE environments. REPHRAIN is fully supportive of the need to protect children online and already has multiple research projects focusing on this area (see also Section 2). However, the centre does not support any of the ongoing arguments for weakening or removing end-to-end encryption in the name of online child protection. The purpose of this evaluation is to provide clear scientific insights into the challenges that need to be addressed when protecting children online within the context of E2EE environments, while also protecting user privacy at scale.

The evaluation process will draw on bimonthly progress reports and technical documents provided by each participating organisation, potentially supplemented by review sessions to answer any additional issues raised by the evaluation team. The evaluation does not include any code review or any form of testing of the proposed solutions within the REPHRAIN centre. Also, since the proposed tools are at the proof-of-concept level, the human rights impact assessment will be limited to the safeguards embedded in their design and those the implementation of which upon deployment is disclosed by the participating organisations. Hence, this work should be interpreted as a useful case study on evaluating (AI-supported) prevention

[2] See e.g. Knodel et al. "Definition of End-to-end Encryption": https://sandbox-ng.ietf.org/doc/draft-knodel-e2ee-definition/and Muffett "A Duck Test for End-to-End Secure Messaging" `https://datatracker.ietf.org/doc/html/draft-muffett-end-to-end-secure-messaging-03`

and detection tools in the context of both sensitive and high-impact online harms (both on the user and potential victim level), while upholding user privacy, security and ethical standards.

The REPHRAIN evaluation is not an endorsement, nor a disapproval of any of the evaluated Proof-of-Concept tools — these are evaluated as exploratory approaches rather than end products. The results of the evaluation process will be made public in a final report to inform future research directions in this area and as a guidance for safety tech industry on how they can further improve and develop their systems.

2 Background of the Evaluation Task

2.1 The Safety Tech Challenge Fund

The Safety Tech Challenge Fund aims to bring together global experts with funding of up to £85,000 each, to demonstrate how end-to-end encryption can be implemented without opening the door to greater levels of child sexual abuse.

The Safety Tech Challenge Fund awarded five organisations from across the world to prototype innovative technologies to help keep children safe in end-to-end encrypted environments, such as online messaging platforms, while ensuring user privacy is respected[3].

Successful applicants are using the funding to develop innovative technologies which demonstrate how tech companies could continue to prevent and detect images or videos showing the sexual abuse of children while ensuring end-to-end encryption is not compromised. Suppliers must demonstrate how their Proof-of-Concept tools protect the privacy of their users, whilst preventing services from being used for the purpose of child sexual abuse.

2.2 The Role of REPHRAIN

REPHRAIN is rooted in an ethos of interdisciplinary research – alongside principles of responsible innovation and creative engagement – to develop new insights that allow the socio-economic benefits of a digital economy to be maximised, whilst minimising online harms that emerge. As such, the centre hosts several experts in Privacy, Security, Artificial Intelligence, Machine Learning, while also leveraging a wide range of socio-technical approaches to online child protection. The research performed in the context of this evaluation underpins REPHRAIN's three core missions, which refer to (1) delivering privacy at scale whilst mitigating its misuse to inflict harms; (2) redressing citizens' rights in transactions in the data-driven economic model by transforming the narrative from privacy as confidentiality only to also include agency, control, transparency and ethical and social values; and (3) addressing the balance between individual agency and social good, developing a rigorous understanding of what privacy represents for different sectors and groups in society (including those hard to reach), the different online harms to which they may be exposed, and the cultural and societal nuances impacting effectiveness of harm-reduction approaches in practice (see also REPHRAIN's scoping document[4]).

REPHRAIN will act as an independent, external evaluator to each of the five projects funded by the 2021 Safety Tech Challenge Fund call to ensure rigour of process and findings can be shared. The proposed Proof-of-Concept tools will be evaluated by a team of REPHRAIN researchers according to strict evaluation criteria, which will include detailed guidance on how the different approaches will need to ensure user privacy.

In this document, the finalised version of the evaluation criteria are published as they will be implemented as part of the REPHRAIN's formal evaluation of each tool. At the end of the five month Delivery Phase the projects will then be evaluated based on the finalised evaluation criteria. Finally, we will publish an evaluation report to share learnings and evaluation results with the community.

[3] More information on these projects can be found on `https://dcms.shorthandstories.com/safety-tech-challenge-fund/index.html`

[4] `https://www.rephrain.ac.uk/scoping-document/`

2.3 REPHRAIN Evaluation Team

The REPHRAIN evaluation team consists of five REPHRAIN researchers with expertise in the field of online child protection, cyber security and privacy, machine learning and artificial intelligence, and socio-technical aspects of human security through developing and applying new technologies:

Claudia Peersman is a Research Fellow at the Bristol Cyber Security Group and one of the core researchers of REPHRAIN. She has been working in the area of developing AI-supported tools for supporting law enforcement investigations pertaining to online harms for over ten years. A key aspect of her research has focused on developing new methods for automatically detecting new or previously unknown child sexual abuse material on P2P networks (iCOP project) and enhancing these techniques to reduce bias towards Western CSAM in current CSAM detection tools (iCOP 2 project [5]). Additionally, she is leading the AUTAPP project[6] (REPHRAIN), in which automated methods are being developed for flagging a range of online harms on social media (e.g. child sexual abuse, exploitation and grooming; cyber-bullying; trolling, aggression and hate speech; depression and self-harm; radicalisation). She is also involved in the ACCEPT project[7] (REPHRAIN), in which she will be investigating the use of PETs and children's rights (e.g. data collection and analysis by smart toys).

Emiliano De Cristofaro is Professor of Security and Privacy Enhancing Technologies at University College London (UCL), where he serves as Head of Information Security Research Group and Director of the Academic Center of Excellence in Cyber Security Research. Emiliano is the co-founder of the International Data-driven Research for Advanced Modeling and Analysis Lab (iDRAMA Lab), sits on the Technology Advisory Panel at the UK Information Commissioner's Office (ICO), and is one of the core researchers, and member of the Leadership Team, of REPHRAIN. His main research interests include problems at the intersection of machine learning and privacy, as well as understanding and countering cybersafety issues using measurement studies and data science. Emiliano's research has been published in several top-tier conferences (IEEE S&P, NDSS, ACM CCS, Usenix Security, WWW, ICWSM, CSCW, ACM IMC, etc.).

Corinne May-Chahal is Professor of Applied Social Science and Co-Director of Security Lancaster, an interdisciplinary ACE CSR and CSE research institute at Lancaster University, and also Chair of the REPHRAIN Ethics Board. Her work involves developing and applying new technologies, with interdisciplinary colleagues, in partnership with industry, the public sector and law enforcement, to address human security in a rapidly changing socio-technical life world. Past projects include; ISIS which created software to identify age and gender deception in computer mediated communication, UDesignIT co-producing applications to facilitate the reporting of community concerns, iCOP (identifying child abuse image originations in Peer to Peer networks), MeSafe (a safeguarding application) and a rapid evidence assessment on victims of online child sexual abuse for the Independent Inquiry into Child Sexual Abuse Internet Investigation. In her latest book Online Child Sexual Victimisation (Policy Press, 2020) she argues for an asset based approach to childhood security; identifying the social assets that are threatened by online harms and developing intersectional strategies on and offline to reinforce these assets (such as the rights to privacy, trust in online services, economic security, freedom of association, freedom from discrimination and violence and promoting wellbeing).

Ryan McConville is a Lecturer in Data Science, Machine Learning and AI at the University of Bristol. His work involves the development of novel machine learning models for large-scale complex data across several modalities. His work is typically applied and evaluated on real world datasets, with interdisciplinary applications in healthcare and cybersecurity. He is leading the CLARITI project [8] in REPHRAIN which is developing multimodal machine learning models to detect online misinformation on social networks by analysing a variety of modalities, including text, images and social behaviour.

[5] https://www.end-violence.org/grants/university-bristol-regional
[6] https://www.rephrain.ac.uk/autapp/
[7] https://www.rephrain.ac.uk/accept/
[8] https://www.rephrain.ac.uk/clariti/

José Tomas Llanos is a Research Fellow at UCL (University College London) Computer Science. Previously, he served as research fellow in Privacy-Aware Cloud Ecosystems (PACE) at UCL's Department of Science, Technology, Engineering and Public Police (STEaPP), and before that as research fellow at the British Institute of International and Comparative Law (BIICL) in the Big data and Market Power project. He has been lecturer in Competition Law at the School of Law of King's College London. He currently acts as consultant for the Organisation for Economic Co-operation and Development (OECD) in matters associated with the digital economy, including privacy, data protection and the economic and social impacts of online platforms. He has experience in interdisciplinary research, having worked with computer scientists to develop a blockchain-based technology capable of enforcing GDPR provisions through smart contracts and flag potential data protection breaches. His research interests revolve around the legal foundations of data protection, the legal status of privacy-enhancing technologies, the implementation of data-protection-by-design principle, and operational gap between law and computer science. His publications focus on competition and big data, the digital economy, data privacy and practical implementation of the GDPR in cloud ecosystems.

3 Revised Evaluation Criteria

The evaluation question is inevitably intertwined with the use of technology to automatically prevent or detect online child sexual abuse material. Up until recently, assessing the performance of such automated tools has generally been based on criteria such as classification accuracy, false positive rates, and usability of the tools. Our work aims to offer a framework for accommodating additional perspectives on evaluating such tools, and how these can be combined. The evaluation criteria are intended to highlight the trade-offs that are faced when selecting different approaches for online child protection purposes in the context of E2EE environments. Additionally, this framework can be applied by the safety tech industry to build public trust in their systems, to positively influence AI technology developments, and to ensure all their users benefit from their solutions.

The initial version of the REPHRAIN evaluation framework includes the following criteria:

1. **Human-centred**. Any system designed to address CSAM should be grounded in human rights[9] and their underpinning values of human dignity and individual autonomy. Any actions performed by the PoC tools that hamper these rights and values, such as deception, unjustified and/or concealed data collection, and discrepancies between the disclosed purpose of the system and the actual actions undertaken by it, are therefore unacceptable. In particular, this criterion focuses on whether and how the PoC tool puts people at its centre, that is to say, it evaluates the manner in which the interests and needs of all its direct and indirect users — i.e., operators, moderators, reviewers, victims and people whose communications are monitored, filtered and/or analysed — are taken into consideration and addressed. This includes, *inter alia*:

    - the comprehensiveness of the PoC tool's functionality, e.g. whether the tool detects only known CSAM (i.e., CSAM already included in existing databases), known and new CSAM, or potentially other types of child abuse, such as violence and online grooming;
    - the implementation of technical, operational and/or organisational measures to avoid the re-victimisation of victims during and after the analysis, and/or to protect the mental health of moderators; and
    - the measures in place to inform people that their communications are screened, blocked and potentially reported, to verify the correctness of the tool's actions (e.g., human review before any content is reported[10]), and to mitigate any potential undue reputational harm or other unfair outcomes (e.g., reports are made only to a competent authority after confirmation of an abuse

[9] See also criterion 2.
[10] See criterion 6.

based on sound predefined criteria, continuous evaluation of machine learning models to rule out bias[11]).

Guiding questions in this regard are: Who are the users of the system and how have they been considered in its design? How do the proposed tools avoid re-victimisation of victims in both existing CSAM databases used by the developed systems and newly detected CSAM? Are CSAM reporting mechanisms (1) included, (2) to whom, (3) triggered under what circumstances? What is the likely impact of the PoC tool on CSAM prevention and the protection of children online more generally?

2. **Human Rights Impact**. To the extent that the PoC tools involve the interception of private communications and/or their metadata to detect, block, investigate and prosecute CSAM online, they may interfere with a number of human rights, safeguards and guarantees enshrined in national laws[12] and international declarations and treaties[13] which the UK is bound to respect and abide by. This criterion is thus intended to assess whether or not the PoC tools have an undue negative impact on:

   2.1. **The Right to Privacy**. The PoC tools must strike an adequate balance between the legitimate aim they pursue — broadly speaking, the protection of children from sexual abuse and exploitation — and the intrusion into the private lives of both users and victims[14] they entail. Thus, the PoC tools must be demonstrably (i) necessary, as opposed to only admissible, ordinary, useful, reasonable or desirable[15], and (ii) proportionate, which involves a rational connection between the tool and aim, as well as the absence of less restrictive means[16]. Fulfilment of these two requirements hinges to a large extent on the scope, extent and intrusiveness of the interference, i.e. on *inter alia*:

   - whether it affects all the users of a service deploying the PoC tool or is targeted to specific users;
   - if targeted to specific users, what elements of suspicion trigger the targeting, including whether or not such determination involves a competent authority;
   - how much personal data and what types of personal data are subject to monitoring, blocking and/or analysis (e.g., images only; images, audio and videos; all the content of communications, including text and metadata, special categories of personal data);
   - whether there is automated decision-making and/or profiling involved (e.g., the automated analysis of text and behavioural patterns to detect potential cases of CSAM dissemination); and
   - whether it is reasonably foreseeable and likely that the PoC tool will be repurposed in the future to detect other types of content, and whether there are technical, operational and legal safeguards to prevent such repurposing.

   Special consideration should be given to the privacy of victims (a vulnerable group), as PoC tools may rely on a machine learning model or a CSAM database which may potentially cause additional harm (e.g., due to bias or unauthorised disclosure). This includes the development stage, as such models require actual CSAM-related data for training and/or testing.

   Guiding Questions: Does the PoC tool imply the general monitoring/scanning/filtering of private communications of all the users of the service implementing it, or does it target specific groups of users? In the last case, which groups, and under what conditions are they targeted?

[11] See criterion 7

[12] Human Rights Act 1998 (HRA), The Privacy and Electronic Communications (EC Directive) Regulations 2003 (PECR), Data Protection Act 2018 (DPA), and the UK GDPR.

[13] See e.g. Universal Declaration of Human Rights (UDHR), International Covenant on Civil and Political Rights (ICCPR), European Convention of Human Rights (ECHR).

[14] User privacy refers to the impact on an E2EE user who would otherwise not have their communications analysed, disseminated or otherwise acted upon. Conversely, victim privacy refers to the impact on a person who appears in CSAM. Occasionally, the user of an E2EE service may also be a CSAM victim.

[15] See e.g., Silver and others v United Kingdom [1983] 5 EHRR 347 at §97

[16] See e.g., Bank Mellat v. HM Treasury (No 2) [2014] AC 700 at §74

Could the PoC tool's aim be achieved through other less-intrusive means? Is the PoC tool likely to be effective in preventing CSAM? Are there any PETs or other safeguards in place to minimise the impact on users' and victims' privacy? What specific types of data will be processed? Is both user and potential victim privacy preserved at different levels: blocking vs. reporting potential CSAM? What is the extent for potential unintended consequences of false positives?

2.2. The Protection of People's Personal Data. Insofar as the PoC tools process personal data, they must observe the data protection principles and safeguards set out in the UK GDPR, the PECR and the DPA. Therefore, PoC tools must at the very least demonstrate compliance with the principles of lawfulness, fairness and transparency, purpose limitation, data minimisation, accuracy, storage limitation, integrity and confidentiality, and accountability[17] (the so-called data quality principles). Furthermore, there must be mechanisms in place to facilitate the exercise of data protection rights[18], and given that the processing at hand is "likely to result in a high risk to the rights and freedoms of individuals", compliance with the aforementioned principles, including the obligation to observe data protection by design and by default[19], should be supported by a data protection impact assessment (DPIA)[20]. The data governance and management plans for all data used and produced by the PoC tools fall within the scope of this evaluation.

Guiding Questions: What is the lawful basis for each type of processing of personal data the PoC tool performs? Is the personal data processed only to detect and block CSAM? What are the technical, operational and organisational safeguards in place to impede the processing of personal data for other purposes? What safeguards have been implemented to comply with the other data quality principles? Is there a draft record of processing activity? How does the PoC tool meet the data protection by design and by default requirements? Has a thorough DPIA been conducted? In what ways and to what extent is the auditability and accountability of the PoC tool ensured[21]? How easy is it for users and victims to exercise their data protection rights?

2.3 The Right to Freedom of Expression. Inasmuch as the PoC tools involve the screening and blocking of messages, images and/or other content an individual intends to send or disseminate to others, they are liable to intrude upon individuals' right to freedom of expression, which includes the freedom to hold opinions, to receive and impart information and ideas, and to access information without undue interference[22]. Just as in the case of the right to privacy, interferences with this right caused by the deployment of the PoC tool must be both necessary and proportionate[23]. Whether or not these requirements are met depends on *inter alia*:

- the extent and scope of the censorship – i.e., the number of people subject to censorship (all the users of the service implementing the PoC tool or specific users) and the type of content subject to screening and blocking (e.g., images only; images, audio and videos; all the content of communications, including text);
- when only specific users are subject to censorship, whether the selection criteria are fair, clear and transparent;
- whether there are sufficient procedural safeguards against the blocking of content[24] – i.e., at the very least notification of the fact that content has been blocked and an appeal process against such action; and

[17] Article 9 UK GDPR.
[18] Chapter 3 UK GDPR.
[19] Article 25 UK GDPR
[20] Article 35 UK GDPR
[21] See criteria 5 and 6.
[22] See Article 10 ECHR. On the importance of access to information and its principle of "free exchange of opinions and ideas" see ECtHR Gillberg v. Sweden, 3 April 2012, § 95, (GC)
[23] It must be noted that censorship prior to publishing is considered the most dangerous, as it stops the transmission of information and ideas to those who wish to receive them. As a result, this type of restriction is subjected to very strict control by the judiciary. See generally ECtHR The Sunday Times v. the United Kingdom (No.2), 26 November 1991 paragraph 51; ECtHR Observer and Guardian v. the United Kingdom, 26 November 1991
[24] See generally ECtHR Cumhuriyet Vakfı and Others v. Turkey, 8 October 2013.

- the availability of remedies for the wrongful removal of content.

Guiding Questions: Does the PoC tool imply the general scanning and blocking of content intended to be sent by all the users of the service implementing it, or does it target specific groups of users? In the last case, which groups, and under what conditions are they targeted? Is the type of content subject to scanning and blocking strictly necessary to achieve the PoC tool's aim? Could the PoC tool's aim be achieved through other less-intrusive means? Is the PoC tool likely to be effective in preventing CSAM? What are the safeguards and redress mechanism in cases of over-censorship or wrongful blocking?

3. **Security**. This criterion aims to ensure that security principles are upheld throughout the lifecycle of each PoC tool. This includes evaluating whether a realistic model that identifies the types of adversaries with an incentive to attack the system (e.g., authorised insiders, outsiders), the most likely adversarial attacks, any potential security vulnerabilities, and what protection mechanisms to address them are in place (e.g., access controls, cryptography, alerts). It also evaluates if proper data and AI/hashing system security measures are in place, and how the CSAM prevention or detection systems are monitored and tested to ensure they continue to meet their intended purpose. Security measures should also include safeguards and mitigation strategies against abuse or unintended use of the systems, especially against wrongful and abusive user reporting (e.g., cryptographic message franking protocols).

   Guiding Questions: Do the PoC Tools have a data diligence process? What security engineering principles and best practices have been observed? What security and mitigation measures are in place regarding potential adversarial attacks, security vulnerabilities and unintended use or abuse of the CSAM prevention or detection systems? How is the PoC Tool's design and implementation verified, validated, tested and monitored?

4. **Effective Performance, Robustness, and Scalability**. An effective and reliable performance is essential in the context of online child protection solutions, both from potential victims' and non-offending users' perspectives. Thus, this criterion focuses on how effective a PoC tool will be in preventing CSAM. This includes analysing how false positives are defined and measured, the implications of the disclosed false positive rate[25], the meaningfulness of evaluation metrics used, the composition of the data used to validate the performance (i.e., the "test data")[26], and what the limitations of each system are. Additionally, it is important to understand a system's robustness to (1) variable non-adversarial circumstances, such as different image or video quality, (2) adversarial behaviour of its users[27], (3) application in different E2EE environments (scalability), and (4) inference in different network conditions or energy levels.

   Guiding questions: Which evaluation metrics are reported? How are false positives defined, measured and reported? What is the false positive rate and what implications stem from it? Are different metrics used for evaluating a system's performance for blocking vs. reporting CSAM? How realistic is the test data used to evaluate the PoC tools? What is the trade-off between the performance rate and the processing time and resources? What are the limitations of each system? How do the solutions perform when applied in different E2EE environments? How do the proposed systems perform under different circumstances (e.g., different quality of video/images, length of videos, embedded CSAM, GIFs)? How do the CSAM prevention or detection systems perform when users attempt to circumvent detection? Do the systems also work offline or on a poor network condition? Is there a trade-off between performance and power consumption of the proposed methods?

5. **Explainability, Transparency, Auditability and Provenance.** The use of automated technologies can have a significant impact on people's lives, especially in the context of online child protec-

[25] The false positive rate can have significant implications for both scalability, user privacy and freedom of expression.

[26] Test data must amount to a realistic set of content, as small differences between the types of content used in evaluation and the types of content shared by E2EE users can lead to significant differences in the false positive rate.

[27] See criterion 3.

tion. Hence, unambiguous justifications for decisions produced by any CSAM prevention or detection system should be available to help users, developers, law enforcement and regulators understand the decision-making process of such tools. This includes reasonable disclosure regarding how and when a CSAM prevention or detection system is engaging with the user, without enabling offenders to circumvent the system. Thus, this criterion focuses on *inter alia*:

- the extent to which the PoC tools, including those based on machine learning models, are auditable — e.g. audits can be performed by anyone or by a trusted third-party only; audits can be made at the source code implementation level or through black-box testing methods; audits may rely on cryptographically verifiable proofs, or on the honesty, skills and diligence of auditing staff only;
- when a tool incorporates data referring to known CSAM content, how is that data audited and authenticated and by whom;
- the manner in which organisations clearly document each step of their pipelines, the development process, testing, limitations, and the intended use of their systems; and
- the degree to which the PoC Tools provide transparency on different levels, e.g. transparency about design, implementation, prior evaluations, training data, matching data, the processes triggered upon CSAM detection, matching results during deployment, false positive rate.

Guiding questions: Do the tools provide an understandable and transparent decision-making process? How do they incorporate the trade-off between responsible disclosures vs. potential adversarial behaviour of offenders? Are the systems' limitations sufficiently communicated and documented? How auditable are the PoC tools, and by whom?; How can machine learning models and known-CSAM databases be audited and authenticated? Do the organisations measure and monitor matching results and the false positive rate, and report on this in a transparent way?

6. **Disputability and Accountability.** Given the potential impact of CSAM prevention or detection tools on a person's human rights, correct system outcomes must be ensured, including on the basis of human oversight and by making available accessible pathways for disputing the decision made by such tools in a timely manner. This includes, *inter alia*, availability of complaint and redress mechanisms in case of wrongful actions (e.g. notification of a blocking decision, and appeal processes against blocking of non-CSAM content) and accountability by the people responsible for different stages of the system's decision-making process.

   Guiding questions: Is human oversight of CSAM prevention or detection tools enabled? Are people responsible for the different stages of the analysis identifiable and accountable for the outcomes of the system? Is there a timely process in place that would allow users to challenge the decisions made by the proposed system?

7. **Fairness/Non-bias.** This criterion aims to ensure that all proposed systems are inclusive throughout their lifecycle. This not only refers to ensuring data diversity during training and testing (e.g. with regard to age group, gender and ethnicity), but also to incorporating fairness metrics into the objective function used to train machine learning models, and to adding constraints into the training process to account for such fairness metrics. Relatedly, this criterion also refers to users receiving equal treatment by the system and equal access to the proposed services.

   Guiding questions: How do the systems perform when applied on CSAM-related data from victims of different age groups, gender and ethnicities? Are debiasing techniques limited to datasets, or do they also involve the system's operation and outputs? Have diverse stakeholder groups been meaningfully involved in the PoC Tool's design?

8. **State-of-the-art.** This criterion evaluates if state-of-the-art research is incorporated in all aspects of the CSAM prevention or detection tools (e.g. children's age detection databases, face recognition when faces are covered).

   Guiding questions: Is the most recent research used to inform the tools? Do the PoC tool's include any innovations building on recent multidisciplinary research?

9. **Maintainability.** This criterion refers to how easily the CSAM prevention or detection tools can be fixed and modified as required. Organisations should have transparent maintenance strategies in place.

   Guiding questions: Are the CSAM prevention or detection tools designed in a way that they can be easily updated, fixed or replaced as required? Are transparent maintenance strategies in place?

4 Acknowledgements

The REPHRAIN evaluation team would like to thank the community again for their time and efforts in supporting our work.