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 and Awais Rashid.

Version 1 - March 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

March 24, 2022

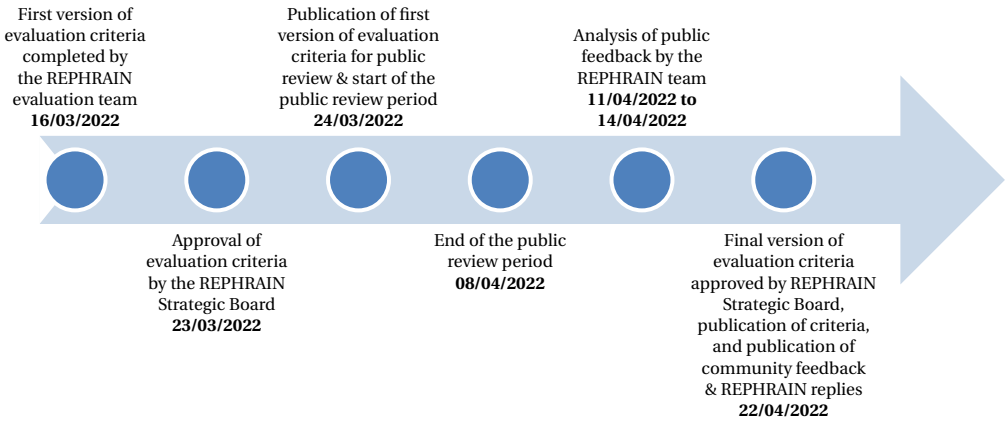
1 Summary

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 different projects funded by the Safety Tech Challenge Fund[1]. 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 publicly publish all results, ensuring that academic rigour and objectivity remain the core of our work, and to inform future directions in this area.

The evaluation criteria that will be developed in the evaluation process are aimed to be a resource for the community and by the community. Hence, with this document, we invite feedback from members of the cyber security & privacy community and stakeholders from academia, industry, law enforcement, and NGOs working in the field of online child protection. The community feedback phase will run for approximately 2 weeks: **from 24 March 2022 until 8 April 2022**.

This formal feedback request will be published on the REPHRAIN website and circulated to the REPHRAIN contact list to ensure maximum exposure. Community feedback and comments can be submitted **online** where all comments will be logged, or can be sent via email to **rephrain-centre@bristol.ac.uk** either as a free form text or an annotated PDF document.

The community feedback will be reported to the REPHRAIN Evaluation Team for full consideration and discussion, and the REPHRAIN Strategic Board will be advised accordingly. Finally, the final version of the evaluation criteria will be published on the REPHRAIN website, along with any major change requests posed by the community, and a discussion on how these requests were addressed, including any changes made (if suitable) and what was the rationale if a change was not made. The timeline for developing the evaluation criteria is as follows:



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

1.1 Objectives and Scope

The REPHRAIN evaluation team aims to provide a **technical assessment** 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 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, any form of testing the proposed solutions within the REPHRAIN centre, or an assessment of the tools for their compliance with legal frameworks on interception of communication, specific AI authorative rules regarding jurisdiction, etc. Hence, this work should be interpreted as a useful case study on evaluating (AI-supported) prevention 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.

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.

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 up 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[2].

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 legitimate users, whilst preventing services from being used by child sexual abuse offenders to facilitate their crimes.

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 the 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[3]).

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, a first draft of the evaluation criteria are published for public review and comment before they are finalised and 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.

2.3 REPHRAIN Evaluation Team

The REPHRAIN evaluation team consists of four 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

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

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

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 [4]). Additionally, she is leading the AUTAPP project[5] (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; cyberbullying; trolling, aggression and hate speech; depression and self-harm; radicalisation). She is also involved in the ACCEPT project[6] (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 [7] 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.

3 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

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

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 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:

Privacy and **Security**. This criterion aims to ensure privacy and data protection principles are upheld throughout the lifecycle of each PoC tool. This includes evaluating data governance and management plans for all data used and produced by the proposed systems. It also evaluates if proper data and AI/hashing system security measures are in place, how potential security vulnerabilities or adversarial attacks can be identified and mitigated, 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 mitigation strategies for abuse or unintended use of the systems.
Example questions: Do solutions have a data diligence process? What mitigation measures are in place regarding potential adversarial attacks, security vulnerabilities and unintended use or abuse of the CSAM prevention or detection systems? Are these systems trained based on data minimisation principles? What PETs are used to protect users'/victims' privacy? Is both user and potential victim privacy preserved at different levels: blocking vs. reporting potential CSAM? What are potential unintended consequences of false positives?

Human-centred. Any system designed to address CSAM should be aligned with human and children's rights, excluding all actions that hamper individual autonomy, such as deception, unjustified data collection, and discrepancies between the disclosed purpose of the system and the actual actions undertaken by the system. Especially in the context of online child protection, this criterion also focuses on how effective a system will be in preventing CSAM, and what measures are in place to avoid re-victimisation of potential victims during and after the analysis.
Example questions: How do the proposed tools avoid re-victimisation of victims in both existing CSAM databases used by the developed systems and newly detected CSAM? Who are the users of the system and how have they been involved in its design? Are CSAM reporting mechanisms (1) included, (2) to whom, (3) likely to be effective? How effective will a system be in preventing CSAM?

Performance, **Robustness**, and **Scalability**. A reliable performance is essential in the context of online child protection solutions, both from a potential victim's perspective and non-offending users' perspective. This includes analysing the meaningfulness of evaluation metrics used, the composition of the data used to validate the performance (i.e. the "test data"), how false positives are defined and measured, 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, and (3) application in different E2EE environments (scalability), and (4) inference in different network conditions or energy levels.
Example questions: Which evaluation metrics are reported? How are false positives defined and measured? Are different metrics used for evaluating a system's performance for blocking vs. reporting CSAM? 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?

Impact on the E2EE ecosystem & access to cryptographic environment. For the purpose of our study, we define end-to-end encryption as the process of encrypting data while it is transferred from one end system or device to another. Hence, the CSAM prevention or detection tools that are deployed before the data is encrypted will not be seen as interfering with E2EE. Likewise, any risks relating to client-side scanning are beyond the scope of this evaluation.
Example questions: Are there any assumptions with respect to keys or access to cryptographic environments? Are intermediary servers used to store data where it can be accessed?

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 users receiving equal treatment by the system and equal access to the proposed services.
Example questions: How do the systems perform when applied on CSAM-related data from victims of different age groups, gender and ethnicities? Have diverse stakeholder groups been meaningfully involved in the design?

Explainability, **Transparency** and **Provenance**. The use of automated technologies can have a significant impact on people's lives, especially in the context of online child protection. 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. Additionally, if a tool incorporates data referring to known CSAM content, it is crucial that such data can be audited and authenticated. Finally, organisations should clearly document each step of their pipelines, the development process, testing, limitations, and the intended use of their systems, to enable rigorous evaluation and help the community to build trust in their proposed solutions.
Example 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? If the system incorporates a blocklist of known CSAM content, how can this be audited and authenticated?

Disputability and **Accountability**. Again, given the potential impact of CSAM prevention or detection tools on a person's life and well-being, efficient monitoring of system outcomes, including human oversight and accessible pathways for disputing the decision made by such tools in a timely manner should be made available. This includes accountability for the people responsible for different stages of the system's decision making process.
Example 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?

Compliance. AI authorative rules are typically issued by different sources for different jurisdictions, and have diverse application domains. This criterion aims to ensure that organisations map out both common AI requirements and requirements that should be addressed in the specific context of online child protection.
Example questions: Did the developing organisations map out common AI requirements that can be assessed as needed (e.g. regarding jurisdiction)? Do the organisations measure and monitor compliance progress and report on this in a transparent way?

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).
Example questions: Is the most recent research used to inform the tools?

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.
Example 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 Community Feedback Request

A community feedback request for changes required to the first version of the evaluation criteria (EC) presented in this document will be open for public review and comment. All constructive comments are welcome, so please let us know what you think. We would appreciate if comments could be based around the following points:

- Positive points on the EC — What did you like about the EC?
- What is missing from the EC and why?
- Should anything be removed from the EC and why?
- How could the EC be improved? Please include examples and references.

Public consultation will be open until **Friday 8 April 2022**.

Comments can be sent by email to **rephrain-centre@bristol.ac.uk** either as a free form text or an annotated PDF document. Comments can also be submitted online.

We would like to thank the community for their time and efforts in supporting our work.