University of BRISTOL

REPHRAIN

Towards a Framework for Evaluating CSAM Prevention & Detection Tools in E2EE Environments: a Case Study

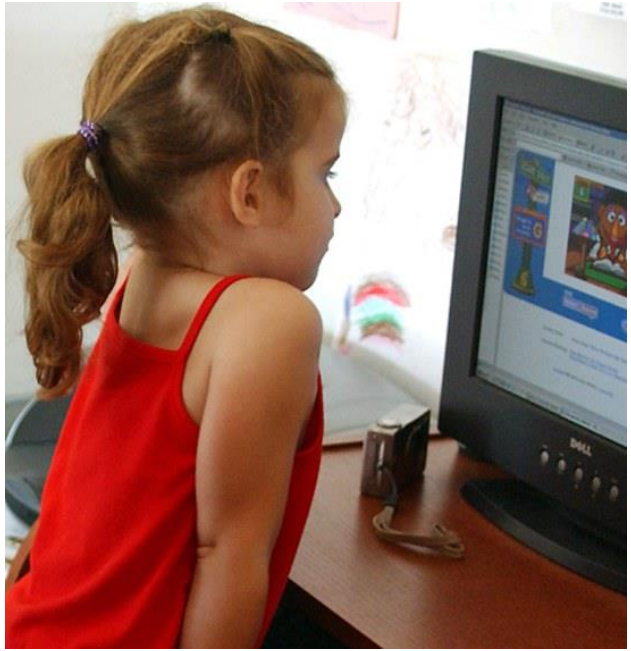
Claudia Peersman
Corinne May-Chahal
Emiliano De Cristofaro
José Tomas Llanos
Ryan McConville
Partha Das Chowdhury
Yvonne Rigby

bristol.ac.uk

REPHRAIN
Protecting citizens online

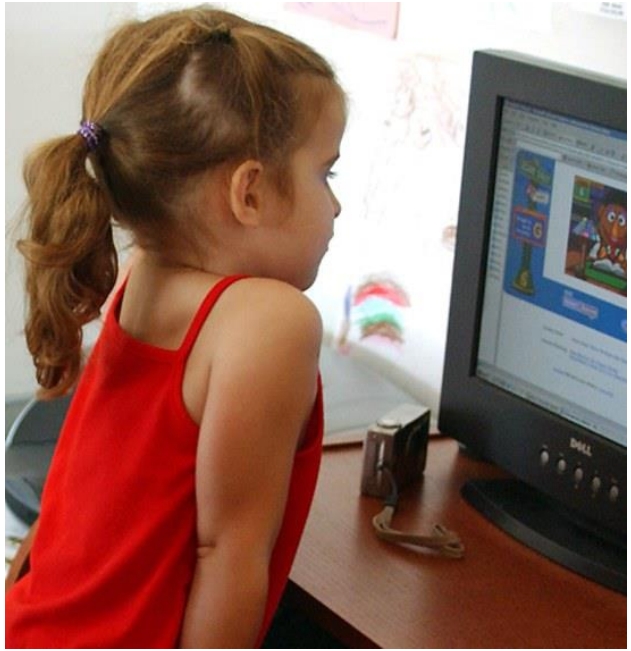
Background

- Safety Tech Challenge Fund, supported by DCMS, HO, ICO and GCHQ
- Focus on development of tools that are able to deploy child safety technologies within E2EE environments, without compromising user privacy
- Detection and/or prevention of child sexual abuse media (CSAM)
- Address specific challenges posed by E2EE environments
- User privacy at the forefront of the approach



Background (2)

- 5 project proposals were selected for funding → 5 months
  - Identify – block – report
  - Known vs new/previously unknown CSAM
  - AI vs hash-based detection
  - General vs full E2EE platform
- REPHRAIN: perform an independent evaluation of their Proof-of-Concept tools



REPHRAIN
Protecting citizens online

The REPHRAIN Evaluation Team

- Online child protection
- Cyber security and privacy
- Machine learning and artificial intelligence
- Legal and human rights aspects in the context of AI







Approach

1. Develop evaluation criteria:
   - Publish draft document with potential criteria
   - Seek input from the community:
     - Cyber security & privacy community
     - Academia
     - Industry
     - Law enforcement
     - Online child protection NGOs
   - Revise the criteria where needed
   - Publish final evaluation criteria

Approach (2)

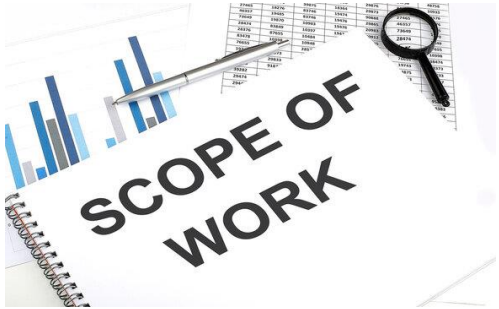
2. Case study: evaluate 5 PoC tools
   - Project descriptions
   - Biweekly progress reports
   - Risk register
   - Review sessions with suppliers

   - **Qualitative analysis →** based on available information
   - Describe the presence or absence of measures that assure compliance with the evaluation criteria

Approach (3)

3. Publish all results:

- Ensure that academic rigour and objectivity remain at the core
- A guidance by safety tech industry to help build public trust in their systems
- Positively influence automated technology developments for online child protection
- Ensure all users benefit from such solutions
- Inform future work in this area



- Technical assessment
- Human rights implications
- Contribute to community debate
  - Privacy vs. online child protection in E2EE environments
  - Trustworthy and ethically responsible AI for online child protection
- **NOT**:
  - Weakening/ “breaking” E2EE
  - Endorsement/disapproval of PoC tools

REPHRAIN
Protecting citizens online

Evaluation Criteria

- Human-centred
- Human Rights impact
- Security
- Effective performance, robustness & scalability
- Explainability, transparency, auditability & provenance
- Disputability & accountability
- Fairness/non-bias
- State-of-the-art
- Maintainability

REPHRAIN
Protecting citizens online

Evaluation Criteria

Human centred

- 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?
- Avoid re-victimisation
- Have users been meaningfully involved in the design?

Evaluation Criteria

Human Rights Impact

- Right to privacy, data protection & freedom of expression
- 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? 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?

Evaluation Criteria

Security

- 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?

Evaluation Criteria

Effective performance, robustness & scalability

- Which evaluation metrics are reported? How are false positives defined, measured and reported?
- 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 are the limitations of each system?
- How do the proposed systems perform under different circumstances?
  - different quality of video/images, length of videos, embedded CSAM, GIFs
  - Different E2EE platforms
  - 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?

Evaluation Criteria

Explainability, Transparency, Auditability and Provenance

- 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?

Evaluation Criteria

Disputability and Accountability

- 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?

Fairness/Non-bias

- 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?

REPHRAIN
Protecting citizens online

Evaluation Criteria

State-of-the-art

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

Maintainability

- 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?

REPHRAIN
Protecting citizens online

Timeline

- Draft evaluation criteria published: 24 March 2022
- Community feedback phase: 24 March – 8 April 2022
- Final evaluation criteria published: 8 July 2022
- Review sessions: September 2022
- Evaluation: October 2022 – January 2023 (delayed)
- Evaluation report published: end of January 2023

Discussion

- Disconnection between academic and private sector research in the field of online child protection

- Focus on improving transparancy and explainability
  → cf. new legislation in Europe, UK whitepaper on online harms

Discussion

- Key challenge: make tools effective & report evaluation without enabling adversarial attacks or compromise commercial interests
- Encourage CSAM detection/prevention tool development from a human lens perspective
- Initial starting point → involve representatives from across the private, public, and civil society, and include children and young people themselves