REPHRAIN
Protecting citizens online

Client-side scanning in private communication: security and privacy risks

Dr Claudia Peersman (University of Bristol), Prof. Corinne May-Chahal (Lancaster University), Dr Partha Das Chowdhury (University of Bristol)

Background

As digital communication platforms like WhatsApp and Messenger increasingly adopt End-to-End Encryption (E2EE) to enhance user privacy, they inadvertently create challenges for detecting and preventing the spread of illegal content, including Child Sexual Abuse Material (CSAM). E2EE ensures that only the sender and receiver can access message content, meaning that no third party, including the platform provider, can intercept or monitor communications. Privacy is essential to human security, freedom of speech, and democracy, but it can also be misused by offenders to evade detection.

To enhance the detection of illegal content like CSAM on these platforms, client-side scanning technologies have been proposed. These technologies scan message content – such as text, images, and videos – on a user's device before the message is sent to the recipient. Content is flagged if it matches or resembles material in a database of illegal content.

However, client-side scanning has drawn significant criticism. In 2023, over 70 researchers raised serious concerns about the UK's Online Safety Bill's proposal for client-side scanning, while over 400 researchers worldwide highlighted similar issues in the EU Draft Child Sexual Abuse Regulation.Their criticisms highlight potential **erosion of privacy, the ineffectiveness of tools due to high rates of false positives and negatives, and the risk of more invasive surveillance.**

REPHRAIN independently evaluated prototype client-side scanning tools developed to detect and prevent CSAM in E2EE environments, funded by the UK Government's Safety Tech Challenge Fund 2021. Resultantly, a comprehensive framework for assessing these tools was developed following an open consultation with stakeholders across academia, industry, law enforcement, and NGOs.

Image credit: Mariia Shalabaieva via Unsplash

Key findings

Our evaluation of the prototype tools in the Safety Tech Challenge Fund 2021 confirmed concerns previously submitted to the UK and European Parliament about the dangers of client-side scanning, namely:

- **Compromising E2EE:** Client-side scanning tools scan content before encryption, undermining the privacy guarantees of E2EE.
- **Ineffectiveness of technology:** Current tools fail to reliably detect CSAM, which risks flagging law-abiding users while allowing offenders to evade detection.
- **Indiscriminate targeting:** Mandatory client-side scanning imposes surveillance on all users, violating privacy rights by treating everyone as a suspect.
- **Risk of 'mission creep':** Client-side scanning could easily expand beyond CSAM detection, enabling mass surveillance of private content for other purposes.
- **Increased security risks:** Adding client-side scanning to E2EE systems creates new vulnerabilities, increasing the likelihood of data breaches, misuse by adversaries, and weakened security for all users.

Ultimately, we found that **none of the tools were fit as a solution to be deployed on E2EE communications.** Along with privacy concerns, our evaluation highlighted:

- **Bias:** Tools were not trained on datasets that included all ethnicities, ages and genders.
- **No guarantee that tools will not be repurposed:** The tools had no built-in safeguards to prevent their repurposing for monitoring personal communications.
- **Transparency and accountability gaps:** There were no clear mechanisms to ensure transparency about who accesses private data or how it is used within client-side scanning systems. Without robust monitoring and oversight from the platforms, it is difficult to ensure that these tools can operate in a way that is transparent and holds developers to account.
- **Testing limitations:** Tools were not tested on actual CSAM due to ethical and practical constraints, leaving gaps in evaluating fairness, robustness, and scalability.

Policy recommendations

Adopt the Evaluation Framework

Our framework has been referenced in policy papers by the OECD and the Directorate-General for Parliamentary Research Services (European Parliament), as well as in research papers.

Policymakers should require tools that aim to detect and prevent CSAM in E2EE environments to be evaluated using this comprehensive framework to ensure they meet ethical and legal standards. Ofcom could peform this function.

Set standards for benchmark datasets

None of the current CSAM detection tools can detect victims of all ethnicities, ages, and genders due to the lack of diversity in benchmark CSAM datasets. Therefore, the government should consult with law enforcement, child protection organisations, the private sector, and researchers to set standards for benchmark datasets to ensure no children are left behind.

Engage children's voices

Throughout the debate around E2EE and how to balance privacy with child protection, children's voices are missing. Children and young people must be engaged in these discussions through various forums including youth advisory panels, workshops, and focus groups.

Listen to expert consensus

Acknowledge concerns from global experts who warn against the risks of client-side scanning in encrypted environments. The government should prioritise these concerns and seek alternative solutions that do not compromise user privacy or security.

Evaluation Framework

This evaluation framework assesses tools to detect and prevent CSAM in E2EE environments against nine key criteria:

1. **Human-Centred Design:** Respecting the rights of all users, including law-abiding individuals, victims, and perceived perpetrators.
2. **Human Rights Impact:** Ensuring compliance with privacy and freedom of expression standards, while minimising harm.
3. **Security:** Ensuring resilience against attacks and misuse without compromising privacy.
4. **Performance:** Effectively detecting CSAM, balancing false positives/negatives, and ensuring scalability.
5. **Transparency:** Making decision-making processes clear, transparent, and auditable.
6. **Accountability:** Providing mechanisms for challenging decisions and holding developers to account.
7. **Fairness and Non-Bias:** Ensuring equitable performance across different demographics.
8. **State-of-the-Art:** Incorporating the latest innovations in CSAM detection.
9. **Maintainability:** Tools can be easily updated and maintained.

The REPHRAIN centre, led by the University of Bristol and partnered with University College London, King's College London, the University of Edinburgh, and the University of Bath, unites experts across disciplines such as Computer Science, Law, Psychology, and Public Policy to explore how to keep people safe online while enabling full participation in digital technologies. Announced by UKRI in October 2020, REPHRAIN now has over 100 experts from 23 UK institutions, working across 50+ research projects to build the UK's leading interdisciplinary community in this mission.

Further information

Access our paper:

Peersman, C. *et al* (2023) 'Towards a Framework for Evaluating CSAM Prevention and Detection Tools in the Context of End-to-end-encryption Environments: a Case Study', *REPHRAIN*. Available at: https://bpb-eu-w2.wpmucdn.com/blogs.bristol.ac.uk/dist/1/670/files/2023/02/Safety-Tech-Challenge-Fund-evaluation-framework-report.pdf.

Contact the researchers:

Dr Claudia Peersman - School of Computer Science, University of Bristol - claudia.peersman@bristol.ac.uk
Prof. Corinne May-Chahal - Department of Sociology, Lancaster University - c.may-chahal@lancaster.ac.uk
Dr Partha Das Chowdhury - School of Computer Science, University of Bristol - partha.daschowdhury@bristol.ac.uk