This article has been accepted for publication in IEEE Security & Privacy. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MSEC.2025.3563092

Weakening end-to-end encryption considered harmful

Awais Rashid, Corinne May-Chahal, Claudia Peersman

Tackling cybercrime and online harms is not a zero sum game. Mitigation measures that may protect one group often open up harms to others. This tension has been most marked in the debates about end-to-end encryption (E2EE) exemplified, over the last 3 years, by the proposed measures regarding scanning of E2EE messaging in the UK Online Safety Bill and the EU Child Sex Abuse Regulation as well as the UK government's notice to Apple to provide access to data stored in iCloud using its Advanced Data Protection product. This has led to serious concerns by civil rights groups, scientists and platform providers regarding the efficacy of such measures and the consequential harms. Multiple open letters from scientists globally have highlighted the flawed assumptions underpinning the mechanisms proposed and the potential risks of mass surveillance and loss of privacy and security due to unfettered access to private (and intimate) communications by platform owners, governments or attackers who may gain access to any systems designed to monitor E2EE communications (https://haddadi.github.io/UKOSBOpenletter.pdf, https://edri.org/wp-content/uploads/2023/07/Open-Letter-CSA-Scientific-community.pdf, https://homes.esat.kuleuven.be/~preneel/Open_letter_CSAR_aug24_still_unacceptable.pdf).

Online child sex abuse and exploitation (OCSEA) is a heinous crime and it is critical that we develop mechanisms that protect the most vulnerable. The harm is not only caused by the crime to create child sex abuse material (CSAM) but continues as this is shared time and again resulting in repeated revictimization. The field has also not been a major focus of research in computer science and security & privacy (S&P), with only a handful of researchers developing tools and techniques to support law enforcement and non-government organizations (NGOs) in tackling the problem. The S&P community needs to engage more with this problem – security & privacy mechanisms can play a key role in safeguarding victims of OCSEA, e.g., through private, secure and safe reporting channels for victims or those aiming to safeguard them.

In this article, we highlight that presenting E2EE as a 'wicked' problem in the context of child protection neither addresses the root causes nor mitigates the harms to the victims. On the contrary, it compromises a key mechanism that can support victims of OCSEA and young people in general in addition to violating the privacy rights of citizens at large. The resulting polarized debates detract from the core need to protect children online and the role that S&P mechanisms can and should play in this regard.

Framing of the problem in technical terms is a non-sequitur

To some, safeguarding victims in E2EE environments is a technical challenge that has the potential to be resolved, if not now, then in the future. But online harm to children is neither caused by E2EE, nor resolvable by its absence. If E2EE did not exist, the problem would remain (as it has done since access to the internet became public). There is an argument that E2EE exacerbates the problem, encouraging abusers to further hide their identities and preventing law enforcement from pursuing

This article has been accepted for publication in IEEE Security & Privacy. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MSEC.2025.3563092

them. But this line of reasoning ignores the decades of CSAM expansion, and the lack of resources to deal with the problem, even before E2EE was generally adopted.

OCSEA is in the main an offline problem, extended and altered by technology. But not by E2EE. The damaging affordances of the internet for abuse are to do with volume, replicability, persistence, manipulability, and audience scale. Most technological developments in the field have focused on improving CSAM detection, as if this is the only way computer scientists are directed to think about the problem. Yet finding more of the massive scale of this content does not solve it.

This automated, large-scale scanning of images and videos shared privately in search of CSAM has been the focus of intense debate due to its potential impact on human rights and fundamental freedoms that are essential to democratic societies. This deeply polarized debate juxtaposes two key arguments: privacy for all users of EE2E environments, arguing that such automated tools are not (yet) fit as a solution versus the protection of children online, suggesting that this type of technology has proven its effectiveness in the, so called, Clearweb applications and law enforcement investigations and its potential impact on Human Rights will be proportional. But how can policy makers know if such tools really work? Up until now, almost all the tools in use (such as those identified by the Bracket Foundation [3]) are commercialized. Whilst they may or may not be effective, evaluations of effectiveness are either not provided or only report on the success of the selected algorithmic approach, focusing on classification accuracy, false positive rates and usability of the tools (cf. [4]). Additionally, policy makers need to rely on the scrutiny of the developers of such tools, given that there has never been an independent, public evaluation of automated industry tools for online child protection.

Tackling abuse is key

OCSEA is gendered. Girls are more likely to be victimized than boys and gender norms in relationships need addressing in the digital world childhood now inhabits. Coercion to create sexual content requires new rules for digital consent and technical processes to support them. There is no doubt that grooming tactics that exploit vulnerability are catalysts for CSAM. If there was no E2EE it might be technically possible to detect more of this, but this has been the case for some time, and again, lack of law enforcement resource has limited effectiveness. The key here is vulnerability. What technical initiatives could reduce vulnerability, help children to recognize vulnerable situations, and increase their resilience? Most online abuse goes unreported. Reporting to service providers is part of the answer, although children are skeptical. What happens because of their report, will it be followed by offers of help? Without answers to those questions, they do not see the point. Can reporting be better scaffolded *securely and privately*, both technically, and socially, within online and offline communities?

Furthermore, technical solutions that do not place victim safeguarding at their core tend to have a displacement effect: abusers move to alternative modes of communication or other platforms to continue with the activity. This has already been observed in this context: as work was done to tackle OCSEA on regular websites, the activity transitioned to peer-to-peer filesharing networks. As more law enforcement activity targeted such networks, the offenders moved to alternative forums

This article has been accepted for publication in IEEE Security & Privacy. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MSEC.2025.3563092

including onion services. We cannot monitor and detect our way out of this problem. We need mechanisms that address the root causes especially suitable reporting mechanisms that safeguard the reporting party's (whether adult or child) privacy and safety from the abuser. E2EE mechanisms are not the problem but a part of the solution in this regard.

Exacerbating harms to privacy

Any mechanism that weakens security and privacy is prone to abuse, whether it is from the developer, those with lawful access to such data and communications (e.g., governments) or attackers who may gain access to such systems. Techniques that are currently in use to detect CSAM, e.g., file hashes can readily be evaded by simple transformations of the image or video. Client-side scanning mechanisms – whereby the communication is scanned on the user's device before being encrypted with an E2EE algorithm – effectively break the fundamental principles of *end-to-end*. This opens up privacy harms, such as risk of mass surveillance, for the population at-large [5]. The counter argument that this would only be done lawfully does not hold. The history of the world is littered with well-intentioned mechanisms by one government that are abused by a subsequent one. Hash values can be replaced by alternatives to monitor, for instance, dissent against government or curtail free speech. Democratic backsliding is a major risk.

Furthermore, the very automation mechanisms on which such client-side scanning relies have limitations in terms of accuracy and false positives. The fact that such false positives are most likely to arise on intimate images and videos shared between consenting adults or innocent family photographs, e.g., of young children playing in baths, leads to serious risk of reputation for individuals and their well-being. We have seen such controversies before, e.g., the outcry over artist Tierney Gearon's images of her children playing on the beach (https://news.bbc.co.uk/1/hi/entertainment/1215944.stm).

With any such system that weakens E2EE, the age old questions and concerns resurface: Who guards the guardians of such systems? How do we ensure that they won't be repurposed or misused by platform providers, governments or other actors who may have access to such systems? How can we assure that attackers would not gain access and compromise privacy of large swathes of the population? There are many high profile examples of breach of government systems or misuse of technologies by democratic governments that should give us serious pause for thought.

The argument isn't only about privacy of adults. Just like adults, children have the right to privacy[1], and they value it.[2] They do not want to lose that right. The downsides of losing online privacy have been well stated. For children, this includes increasing the potential for mass commercial exploitation and playing out their childhood under greater surveillance than any other previous

[1] Article 16 United Nations 1989 Convention on the Rights of the Child. https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-rights-child

[2] What do children want to help them stay safe online? Views from Children in Blackpool. https://youtu.be/n6Ugzke_9GY

This article has been accepted for publication in IEEE Security & Privacy. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MSEC.2025.3563092

generation. Finding workarounds such as client-side scanning and other technical solutions to the E2EE challenge is not the answer. The problem demands whole system frames and solutions that are evidence based, co-designed with children and parents, civil society institutions, business and service providers.

Policy makers should understand that there are currently no straightforward solutions to these challenges. Our recent work at the UK National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online (https://www.rephrain.ac.uk/), in which we analyzed five Proof-of-Concept tools designed for CSAM detection in E2EE environments has shown that striking a fair balance amongst the rights and interests of all individuals concerned (law-abiding users, CSAM victims and perceived perpetrators) proved to be a key issue. Although none of the proof-of-concept tools proposed to weaken or break the end-to-end encryption protocol, the confidentiality of the E2EE service users' communications could not be guaranteed. All content intended to be sent privately by every user of the E2EE service is monitored pre-encryption, in such a way that everyone is treated as a potential suspect of CSAM-related crimes, and in some cases could be collected for training/fine-tuning machine learning models.

AI is no silver bullet

The advances in machine learning, deep learning, and more recently, generative artificial intelligence have sharpened the focus on the potential of such techniques in detecting and apprehending those engaged in sharing of CSAM. However, as noted above, there is a lack of systematic and independent evaluation of such techniques this context. Additionally, due to the lack of diverse benchmark datasets for developing and evaluating on-line child protection tools, there is no way of guaranteeing that the tools would be able to detect victims of all ethnicities, age and gender groups.

Evaluation is important but just as concerning are the unintended consequences of AI. This includes AI generated CSAM content, which has led the UK government to introduce the first AI related offences in the new Crime and Policing Bill (https://www.gov.uk/government/news/britains-leading-the-way-protecting-children-from-online-predators) and AI generated algorithmic drivers encouraging internet users to view CSAM content [7].

Returning to our argument that a focus on E2EE detracts from the core of the problem, we note that it is critical that policy makers focus on establishing (1) an agreed (international) human-centric framework that engages with the fundamental issues at play and takes into consideration the impact of any technological mechanisms on the privacy and human rights of citizens, (2) diverse and ethically responsible benchmark datasets, enabling an independent assessment of the effectiveness any automated tools for OCSEA detection and mitigation and (3) concerted efforts to advance socio-technical initiatives that reduce vulnerabilities to technology facilitated and enabled harms, including OCSEA, and increase human security and resilience.

This article has been accepted for publication in IEEE Security & Privacy. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MSEC.2025.3563092

In conclusion

The key to breaking the deadlock in this debate is to focus on the collective goal: protecting children from abuse while also ensuring Human Rights are respected and end-to-end-encryption is not compromised. The two objectives should not be at odds in any solutions we devise. The security & privacy community has a key role to play – both in terms of protecting privacy of users of online services and in developing mechanisms that enable privacy and safeguarding of those working to support victims of OCSEA and the victims themselves.

References

[1] Kardefelt-Winther, D., Day, E., Berman, G., Witting, S.K., and Bose, A., on behalf of UNICEF's crossdivisional task force on child online protection (2020). Encryption, Privacy and Children's Right to Protection from Harm. Innocenti Working Paper 2020-14. Florence: UNICEF Office of Research – Innocenti.

[2] Corinne May-Chahal, Claudia Peersman, Awais Rashid, Maggie Brennan, Emma Mills, Peidong Mai, John Barbrook (2022). A Rapid Evidence Assessment of Technical Tools for the Detection and Disruption of Child Sexual Abuse Media (CSAM) and CSAM Offenders in the ASEAN Region. Technical Report. University of Bristol. https://research-information.bris.ac.uk/en/publications/a-rapid-evidence-assessment-of-technical-tools-for-the-detection-

[3] Bracket Foundation (2019). Artificial Intelligence: Combating Online Sexual Abuse Of Children, https://respect.international/wp-content/uploads/2019/11/AI-Combating-online-sexual-abuse-of-children-Bracket-Foundation-2019.pdf

[4] Muhammad Uzair Tariq, Afsaneh Razi, Karla A. Badillo-Urquiola, Pamela J. Wisniewski (2019). A Review of the Gaps and Opportunities of Nudity and Skin Detection Algorithmic Research for the Purpose of Combating Adolescent Sexting Behaviors. HCI (3): 90-108

[5] Harold Abelson, Ross J. Anderson, Steven M. Bellovin, Josh Benaloh, Matt Blaze, Jon Callas, Whitfield Diffie, Susan Landau, Peter G. Neumann, Ronald L. Rivest, Jeffrey I. Schiller, Bruce Schneier, Vanessa Teague, Carmela Troncoso (2024). Bugs in our pockets: the risks of client-side scanning. J. Cybersecur. 10(1).

[6] Claudia Peersman, José Tomas Llanos, Corinne May-Chahal, Ryan McConville, Partha Das Chowdhury, Emiliano De Cristofaro (2023). "Towards a framework for evaluating csam prevention and detection tools in the context of end-to-end encryption environments: a case study," https://www.rephrain.ac.uk/safety-tech-challenge-fund/.

[7] see Insoll, T., Soloveva, V., Díaz Bethencourt, E., Nieminen, N., Leivo, K., Ovaska, A., & Vaaranen-Valkonen, N. (2024). What Drives Online Child Sexual Abuse Offending? Understanding Motivations, Facilitators, Situational Factors, and Barriers. 2KNOW project. https://www.suojellaanlapsia.fi/en/2know-research-report.

About the authors

This article has been accepted for publication in IEEE Security & Privacy. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MSEC.2025.3563092

Awais Rashid is Professor of Cyber Security at University of Bristol, UK. He is Director of the National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online (REPHRAIN). Contact him at awais.rashid@bristol.ac.uk

Professor Corinne May-Chahal is Professor of Applied Social Science at Lancaster University. She leads Co-Designing Community Resilience to Online Child Sexual Victimisation in the Vulnerability and Policing Futures ESRC Research Centre (ES/W002248/1). https://vulnerabilitypolicing.org.uk/online-child-sexual-victimisation/. Contact her at: c.may-chahal@lancaster.ac.uk.

Dr. Claudia Peersman is a Research Fellow at the University of Bristol's Cyber Security Group and one of the core researchers of the UK National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online (REPHRAIN). Her research spans text mining and cyber security and focuses on developing new, AI-supported tools and techniques to support law enforcement agencies in their investigations pertaining to cyber crime. More particularly, her work has focused on automatically detecting CSAM on P2P networks, identifying grooming and deceptive users in online social media and analysing cybercriminal communications on Darknet markets. A key aspect of her research has focused on enhancing CSAM detection techniques to reduce bias towards Western CSAM and on defining trustworthy and human-centred AI by investigating if and how existing guidelines can be tailored to the highly sensitive context of online child protection. Contact her at: claudia.peersman@bristol.ac.uk.