REPHRAIN

Protecting citizens online

REPHRAIN:

Towards a Framework for Evaluating CSAM Prevention and Detection Tools in the Context of End-to-end encryption Environments: a Case Study

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

Version 1 - February 2023

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

Towards a Framework for Evaluating CSAM Prevention and Detection Tools in the Context of End-to-end-encryption Environments: a Case Study

Version 1

February 23, 2023

Executive Summary

This report focuses on the development of a framework for evaluating automated CSAM detection and prevention tools in the context of End-to-End Encryption environments. Given the tensions that arise between protecting vulnerable users like children and safeguarding privacy and security at large on such platforms, we discuss a set of criteria that go beyond classification accuracy, false positive rates, and usability. We present a human-centred framework, supported by the research community that also incorporates Human Rights implications, security, explainability, transparency, fairness, accountability, disputability, relation to the state of the art, and maintainability. We aim to contribute to ongoing research 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 in E2EE environments.

We also report on a case study whereby the evaluation framework was implemented as part of an independent formal evaluation of five Proof-of-Concept (PoC) tools funded by the Safety Tech Challenge Fund. We discuss the challenges arising when assessing industry tools for online child protection without access to commercially sensitive information. To our knowledge, this is the first independent, public evaluation of automated industry tools for CSAM detection and prevention.

Given the recent regulatory focus both in the UK and Europe on transparency and explainability in the use of automated tools for online child protection, our evaluation framework is meant to: (1) provide guidance for the safety tech industry on how they can further improve and develop human centred, trustworthy systems for online child protection, and (2) inform different stakeholders such as policymakers, law enforcement, and researchers about key challenges and limitations.

Despite the exploratory nature (i.e., "the art of the possible") of the PoC tools evaluated and the evaluation criteria being published post-hoc, this report highlights important questions that must be addressed when building online child protection tools, especially in the highly complex context of E2EE environments.

Overall, the main takeaways of the report can be summarised as follows:

1. Striking a fair balance between the rights and interests of all individuals concerned, i.e., law-abiding users, (potential) CSAM victims, and perceived perpetrators, is a key issue. Although none of the PoC tools propose to weaken or break the end-to-end encryption protocol, from a Human Rights perspective, the confidentiality of the E2EE service users' communications cannot be guaranteed when *all* content intended to be sent privately within the E2EE service is monitored pre-encryption. This contrasts with online child protection tools currently deployed on non-private online communications.

2. On the one hand, designing a CSAM detection/prevention tool with the potential of easily re-purposing the technology to detect/prevent other types of illegal or unwanted content, or potentially collecting users' communications for retraining/fine-tuning machine learning models seems valuable from a commercial point of view. On the other hand, however, it is highly concerning in the context of analysing protected communications. Therefore, we argue that it is essential to include technical, legal, operational, and/or contractual safeguards *by design* to prevent the re-purposing of such technologies prior to any deployment in a real-life E2EE application.

3. Meeting the transparency, disputability, and accountability criteria proved to be difficult. Despite the products are in an early development stage, we strongly advise these principles to be considered by design, rather than relying on the scrutiny of the E2EE platforms into which they might be integrated.

4. Arguably, the biggest challenges in applying the evaluation framework stems from the absence of: (1) documented, ethically responsible benchmark datasets for developing and evaluating CSAM detection/prevention tools and (2) detailed experimental information due to confidentiality issues. As

a result, none of the PoC tools could be assessed for their fairness/non-bias, performance, use of state-of-the-art techniques, robustness, or scalability. Establishing benchmark datasets would enable the independent evaluation of online child protection technologies without the risk of compromising commercial interests.

1 Introduction

The continued growth in child sexual abuse material (CSAM), markedly during the pandemic (e.g., [8]), suggests that despite all efforts — technical and social — the prevention of CSAM is a major challenge. CSAM derives from several sources [16], including:

- Peer on peer coercive image sharing, originating in schools, gangs and within offline peer relationships;
- Grooming online that leads (or intends to lead) to offline contact involving the production and sharing of images and videos as part of the grooming process;
- Peer to peer exchange and communication between offenders using a variety of platforms (e.g., peer-to-peer (P2P) networks, end-to-end encryption (E2EE) environments, dark web fora), depicting CSAM produced in offline settings including recorded sexual abuse of children (particularly young children) in domestic and childcare environments;
- Live streaming of child sexual abuse in a commercial (often international) context, where the abuse is committed in offline settings and engaged with virtually by perpetrators, then capped for future viewing;
- Viral image sharing, where CSAM is shared "in disgust or misplaced humour" [13], leading to revictimisation of the child depicted and contributing to many of the reports to clearing houses.

Each of these sources requires preventative efforts at technical, law enforcement and civil society levels. An understanding of the different online manifestations of CSAM is critical to effective tool development.

In 2021, the Safety Tech Challenge Fund awarded funding to five projects to prototype innovative, automated technologies to help keep children safe in E2EE environments, such as online messaging platforms, while ensuring user privacy is respected. As is standard with innovation funding, the Safety Tech Challenge Fund's purpose was not to develop fully-fledged tools that were ready for implementation in a commercial setting, but to support the development of Proof of Concept (PoC) technologies that test the art of the possible (see also Section 2.1). To enable academic scrutiny, the UK National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online (REPHRAIN) was requested to act as an independent, external evaluator to each of these five projects (see Section 2.2). To our knowledge, this is the first independent, public evaluation of automated industry tools developed for online child protection.

Up until recently, assessing the performance of automated tools in this field has generally been based on criteria such as classification accuracy, false positive rates, and usability of the tools [24]. Given the recent regulatory focus both in the UK and in Europe on transparency and explainability in the use of automated tools for online child protection (e.g., [7, 5], this work aims to contribute to the development of a framework for accommodating additional perspectives on evaluating such tools, and how these can be combined. More specifically, in this study we present:

- the finalised version of the evaluation criteria, which aside from performance also include criteria focusing on human rights impact, security, explainability, transparency, fairness, accountability, etc. The 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. Addition-

ally, they can be used as a guidance by the safety tech industry to help build public trust in their systems, to positively influence AI technology developments for online child protection, and to ensure all users benefit from these solutions (Section 4.1).

- a case study in which these criteria were implemented as part of a formal evaluation of each Proof-of-Concept (PoC) tool funded by the Safety Tech Challenge Fund. We describe the presence or absence of different measures that assure compliance with each criterion and provide guidance where possible when certain criteria were not met (Section 5).
- a discussion on how future research can support a framework for evaluating CSAM detection and prevention tools (Section 7).

This study aims to contribute to recent work defining trustworthy and human-centred AI (see for example [14, 10, 24]) by investigating if and how existing guidelines can be tailored to the highly sensitive context of online child protection in E2EE environments. Research with a specific focus on challenges when developing trustworthy AI (or automated techniques) for detecting and preventing sensitive, high-impact online harms is currently still limited in the field.

Additionally, this work discusses the challenges that arise when assessing industry tools for online child protection without potentially compromising their commercial interests. As discussed in Section 5, the information about each PoC tool that was made available to the team was insufficient to evaluate their compliance to some of the evaluation criteria, such as performance, fairness/non-bias, robustness and scalability.

Finally, it is important to note that this study focuses on a case study which includes a static view of the evaluated tools. More specifically, the REPHRAIN team analysed the projects' progress reports submitted to the Safety Tech Challenge Fund between December 2021 and April 2022, along with each project's description, risk register and project plan. Given the exploratory nature of each PoC tool and the publication of the evaluation criteria in April 2022, the findings presented in this report may not transfer to the current status of these tools' development. Hence, we emphasise that this study does not provide an endorsement, nor a disapproval of any of the evaluated tools.

2 Background of the Evaluation Task

2.1 The Safety Tech Challenge Fund

As children today continue to face a sustained threat of sexual exploitation and abuse online, with COVID-19 further exacerbating this risk, moving quickly to scale up and speed up the development of vital online safety technology solutions to actively prevent abuse and reduce the proliferation of child sexual abuse material on digital platforms is crucial.

This vision rests on the collaboration between governments, private sector companies, NGOs, and academics to develop innovative, responsible technologies to tackle harmful and illegal behaviours taking place on social media and other online platforms. The Safety Tech Challenge Fund (STCF) was established to execute this task, whilst ensuring end-to-end encryption is not compromised.

Through the Safety Tech Challenge Fund, five projects were awarded an initial £85,000 each in 2021, with a further £129,500 in stretch funding later shared between two of the five projects (Cyacomb and DragonflAI, see below) to prototype innovative ways in which sexually abusive images or videos of children can be detected and addressed within E2EE encrypted environments, while ensuring user privacy is respected. The five funded projects were: Cyacomb Safety, SafeToWatch, GalaxKey, DragonflAI, and T3K. More details about each tool can be found in Section 5. The Fund has provided these project suppliers with a five-month prototype building phase, during which they received mentorship and support from DCMS, the Home Office,

GCHQ, ICO, and delivery partner, PUBLIC.

In the final stage, each resulting PoC tool was evaluated by a team of REPHRAIN researchers (see below), who performed their evaluation task independently from the Fund and any of the other organisations mentioned above. The latter goes beyond the competition brief provided to successful bidders through the Challenge Fund (see [4]).

2.2 REPHRAIN

2.2.1 Role

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[1]).

REPHRAIN has accepted the request to act as an independent, external evaluator to each of the five projects (see Section 2.1) funded by the 2021 Safety Tech Challenge Fund call to ensure rigour of process and findings can be shared. A team of REPHRAIN researchers (see Section 2.2.2) have drafted a set of draft evaluation criteria, which were published for public feedback. In this document, the finalised version of the evaluation criteria are published (see Section 4.1) as they were implemented as part of the REPHRAIN's formal evaluation of each tool. Additionally, the results of the evaluation are discussed in Section 5.

2.2.2 Evaluation Team

The REPHRAIN evaluation team consists of six 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 [2]). Additionally, she is leading the AUTAPP project[3] (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[4] (REPHRAIN), in which she is investigating the use of PETs and children's rights (e.g. data collection and analysis by smart toys).

[1] https://www.rephrain.ac.uk/scoping-document/
[2] https://www.end-violence.org/grants/university-bristol-regional
[3] https://www.rephrain.ac.uk/autapp/
[4] https://www.rephrain.ac.uk/accept/

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).

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.

Ryan McConville is a Senior 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 to, and evaluated on, real world datasets, with interdisciplinary applications in healthcare and cybersecurity. He is leading the CLARITI project [5] 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. His research has been published in some of the most prestigious venues.

Partha Das Chowdhury is a Research Associate at REPHRAIN (University of Bristol, UK). His research interests are in privacy enhancing technologies, secure software development, security protocols and adoption of tools to protect citizens from online harm. He is currently leading the development of the REPHRAIN privacy testbed. He has experience in evaluating E2EE desktop clients along with researchers from the University of Cambridge. Partha brings key insights from other disciplines and industry to shape his research. He was the first to propose the use of Capability Approach as the foundation of designing protection mechanisms. The paper was published at NSPW, 2022. He has over a decade of industrial technology implementation experience for various client organisations including Tata Steel, city traffic management systems and mining majors. He was associated with the Center for 4th Industrial revolution, an initiative of the World Economic Forum as one of the contributors for the Blockchain toolkit. He was invited as an expert to the Commonwealth Working Group on Interception of Communication and Related Matters at Marlboro House, London in 2005.

[5] https://www.rephrain.ac.uk/clariti/

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), 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.).

Additionally, this document was reviewed and approved by REPHRAIN's Strategic Board prior to publication.

3 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 (PoC) 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 E2EE 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[6]. 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, broadly conceived. 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 task 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 drew on bimonthly progress reports and technical documents provided by each participating organisation, supplemented by a review session with each project to address 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 was limited to the safeguards embedded in their design and those the implementation of which upon deployment was disclosed by the participating organisations. Hence, this work should be interpreted as a first step towards evaluating automated prevention and detection (industry) tools in the context of 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 are made public in this final report to inform future research directions in this area and as guidance for the safety tech industry on how they can further improve and develop their systems.

[6] 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`

4 Approach

Key steps in REPHRAIN's evaluation process were to (1) develop a draft list of potential evaluation criteria, (2) seek input from the community and revise the criteria where needed, (3) assess the five PoC tools based on the finalised version of the evaluation criteria, and (4) publish all results, ensuring that academic rigour and objectivity remain at the core of our work, and they can inform future work in this area. This section describes each step of our methodology.

4.1 Evaluation Criteria

The evaluation criteria developed by the REPHRAIN evaluation team aim to be a resource for the community and by the community. Hence, during the scoping stage of the task, 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 were published on the REPHRAIN website (cf. here), along with a summary of the key changes made. It is important to note that these evaluation criteria were developed independently of the Safety Tech Challenge Fund without oversight or input from the funders or suppliers.

The final version of the REPHRAIN evaluation framework includes the following criteria[7]:

1. **Human-centred**. Any system designed to address CSAM should be grounded in human rights[8] 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[9]), and to mitigate any potential undue reputational harm or other unfair

[7] The evaluation criteria are repeated here for the readers' convenience.
[8] See also criterion 2.
[9] See criterion 6.

outcomes (e.g., reports are made only to a competent authority after confirmation of an abuse based on sound predefined criteria, continuous evaluation of machine learning models to rule out bias[10]).

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[11] and international declarations and treaties[12] 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[13] they entail. Thus, the PoC tools must be demonstrably (i) necessary, as opposed to only admissible, ordinary, useful, reasonable or desirable[14], and (ii) proportionate, which involves a rational connection between the tool and aim, as well as the absence of less restrictive means[15]. 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

[10] See criterion 7

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

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

[13] 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.

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

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

groups of users? In the last case, which groups, and under what conditions are they targeted? 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[16] (the so-called data quality principles). Furthermore, there must be mechanisms in place to facilitate the exercise of data protection rights[17], 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[18], should be supported by a data protection impact assessment (DPIA)[19]. 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[20]? 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[21]. 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[22]. 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[23] – i.e., at the very least notification of the fact that content has been blocked and an appeal process against such action; and

[16] Article 9 UK GDPR.
[17] Chapter 3 UK GDPR.
[18] Article 25 UK GDPR
[19] Article 35 UK GDPR
[20] See criteria 5 and 6.
[21] 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)
[22] 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
[23] 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 mechanisms 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 whether 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 in E2EE environments. This includes analysing how false positives are defined and measured, the implications of the disclosed false positive rate[24], the meaningfulness of evaluation metrics used, the composition of the data used to validate the performance (i.e., the "test data")[25], 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[26], (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 in 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-

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

[25] 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.

[26] 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?

In the following section, we discuss how each PoC tool was evaluated based on these criteria.

4.2 Evaluation Method

Between December 2021 and April 2022, suppliers submitted bimonthly progress reports to the STCF delivery partner PUBLIC. These reports were made available to the REPHRAIN evaluation team, along with each project's description, risk register and project plan. As the final version of the evaluation criteria was developed and published post April 2022, and hence post the PoC tools' development stage and delivery of the final progress reports, a review session was set up between the team and each supplier in September/October 2022 to highlight what additional information was needed to enable a more complete evaluation of each PoC tool according to these criteria.

Given the exploratory nature of each PoC tool — in contrast to tools that are said to be ready for deployment — it was agreed not to include any type of scoring in our evaluation. Instead, assessment was based on a qualitative analysis of each tool, which was restricted to the information that was made available by each supplier. Due to confidentiality issues, we were not provided with detailed experimental results for any of the PoC tools. The evaluation team decided not to enter into confidentiality agreements where they were offered by suppliers to obtain such information, as that would have impacted the team's commitment to providing transparency of the evaluation results. Finally, the STCF and the suppliers were provided with 48 hours for checking factual errors prior to publication of this report. Of all comments received, only those that related to factual errors were addressed by the team.

In the next section, we describe the presence (or absence) of different measures that assure compliance with the evaluation criteria and provide guidance where possible when certain criteria were not met.

Given the limitations mentioned above, i.e. the post hoc publication of the evaluation criteria, the exploratory nature of the tools and the limited information available at the time of the evaluation, we emphasise that the evaluation presented in Section 5 does not provide an endorsement, nor a disapproval of any of the evaluated tools. By highlighting which (aspects of the) criteria were not met, this case study shows how the evaluation framework presented in this work can be used to inform their next stage of development.

5 Evaluation Results

5.1 Project Cyacomb Safety

Cyacomb Safety is supplied by Cyacomb, Crisp Thinking, University of Edinburgh, and the Internet Watch Foundation (IWF). The Cyacomb Safety project's aim was to enhance (1) Cyacomb's new 'Contraband Filter' technology (a database built from databases of harmful or illegal content) and (2) their new type of Privacy Assured Matching protocol, which works using the Contraband Filter to enable split matching (where a device or client and server cooperate to determine matching to a Contraband Filter) and is targeted at E2EE messaging applications, on social media and in cloud platforms to detect known CSAM and terrorist content. More specifically, their PoC tool allows for an exchange of data between an application on the user device (e.g. at application or operating system level) and a cloud platform, thereby potentially triggering a match with a CSAM/terrorist database in the form of a Contraband Filter. If contraband content is detected, the user's file will not be sent.

When analysing the tool's compliance to our **Human-centred** criterion, it is key to include the rights and interests of all individuals concerned — i.e. law-abiding users, perceived perpetrators and potential CSAM victims. In this regard, Cyacomb claims that the Privacy Assured Matching protocol is designed to ensure that an attacker with access to the communication between the device and the server can only access a minimum amount of information about the content being examined. This information should be insufficient to make any meaningful guess about the nature of such content, and therefore the confidentiality of the E2EE service users' communications should be preserved, at least against attackers snooping the network or seeking to intercept communications. In turn, since it aims to identify and block known CSAM before it enters an E2EE environment (reporting of CSAM being optional), the impact on perceived perpetrators' fundamental rights to privacy and freedom of speech might be justified to some extent. However, safeguards would be needed to ensure that the scope of such analysis cannot be widened. We further discuss relevant issues pertaining to scanning scope below. Lastly, and crucially, technical and operational measures were reported to be implemented to prevent the re-victimisation of CSAM victims by ensuring that known CSAM is not further shared or leaked. More specifically, Cyacomb partnered with the Internet Watch Foundation (IWF) to produce a CSAM Contraband Filter and test Cyacomb's solution with actual CSAM data within the IWF's secured environment. This testing could be carried out without anyone outside IWF being exposed to CSAM, and without any CSAM or CSAM-related data leaving the IWF's environment. Moreover, they claim that the Privacy Assured Matching protocol is designed to ensure that an attacker intercepting the device-server communication cannot determine whether a match was made, or even to definitively match the content if they have other data suggesting that it may be.

Encryption technologies contribute in a fundamental way to the respect for private life and confidentiality of communications, the protection of personal data, freedom of expression, and the protection of democracy in general. Any interference with these rights must therefore be strictly necessary and proportionate for achieving the intended objective of protecting children against sexual harm. As noted in the evaluation criteria, whether these requirements are met depends on the scope and extent of the measure, its level of intrusiveness, whether one or more **fundamental rights** are encroached upon, and the existence of safeguards against errors and abuse.

Cyacomb's PoC tool involves the automated filtering and scanning of all the relevant E2EE service's users' content data, with a view to detecting known CSAM images and videos. As a result, the scope and extent of the PoC tool is undoubtedly broad: according to well-settled principles of international human rights law (e.g., ECrtHR (Karabeyoğlu v. Turkey, § 103), any analysis of private content should target only people that are under investigation based on specific, reasonable and individual-level suspicion, and not other users of the relevant service. Nonetheless, since the automated analysis of users' content is performed to match images and videos on the basis of previously confirmed instances of CSAM, the PoC tool's level of intrusiveness is lower than that arising from the processing of content data to detect new CSAM or child grooming, which typically involve the automated analysis of non-CSAM content, text and speech. Put in other words, it is the *least intrusive* measure as compared to other technologies that rely on automated processing and AI to detect child sexual abuse material. Moreover, the PoC tool's reliance on cryptographic hashing of confirmed CSAM ensures it achieves its intended aim.

However, the threat to freedom of speech posed by Cyacomb's PoC tool cannot be overlooked. Broadly speaking, automated filtering and scanning of all users' content data at the moment of dissemination can be understood as a form of prior restraint, or prior censorship, against the validity of which there is a strong presumption in human rights law. This is because, instead of enduring preventive scanning and filtering of their speech, people must be free to speak their minds, and then face potential punishment should their speech amounts to a law violation. Most importantly, as Apple's recent proposal for a cryptography-based privacy-preserving CSAM detection system showed, CSAM detection systems, such as Cyacomb's PoC tool, are inherently dual-use technologies. Thus, there is a risk that, due to changes in regulations, policy, or demands from foreign governments, the same methods Cyacomb uses for identifying CSAM could be applied to other content that amounts to free speech (see e.g., [11]). In the case at hand, this risk is very real, as Cyacomb observes that its Contraband filter technology is also used in counter-terrorism and online safety applications. Technical, contractual, legal and operational safeguards should be developed against the re-

purposing of the PoC tool to monitor and detect other content. Otherwise, deployment of the PoC tool can be neither necessary nor proportional from a human rights point of view.

Another crucial factor to assess the proportionality of the restrictive measure (i.e. the PoC tool) is the availability of safeguards against errors or unfair outcomes. Since the potential reporting of CSAM to the competent authorities (if the reporting functionality is enabled) after automated detection may significantly affect the data subject concerned, no report should be based solely on the outcome of automated detection, not least where there is room for false positives. Otherwise, individuals' rights under Article 22 UKGDPR would be breached. An optional feature of Cyacomb's Privacy Assured Matching is to make reports readable only when they are 'true-positive' reports; reports arising from false positives in Cyacomb's system are thus unreadable. However, it is unclear from the available documentation how the 'true-positive' quality of a report is ascertained (e.g. based on human review or statistical accuracy), and at any rate, there does not seem to be any safeguard against reporting false positives, even in unreadable form. In addition, due process and freedom of speech considerations dictate that, in addition to data protection safeguards, there must be effective mechanisms and remedies in place for cases where user content has been blocked or removed, or a user's content or identity have been reported to competent authorities in error, such as notifications and an appeal process. The PoC tool contemplated no measures of this kind at the point of evaluation.

For matching purposes, Cyacomb's Contraband Filter technology's **performance** is said to be comparable to using cryptographic hashes. The tool is reported to "highly reliably detect" content that is an exact match for the database the Contraband Filter was built from. However, no evaluation information was shared with the REPHRAIN team. Cyacomb did report that variations of content (e.g. different quality, adding noise, cropping) will prevent matches. False positive rates were defined as the equivalent of hash collisions, but it is unclear how they are measured. Test data was obtained through collaboration with IWF, but lack of key experimental detail impedes a rigorous assessment of the tool's performance, robustness and scalability. Cyacomb suggested the use of perceptual or similarity matching techniques in their reports to deal with some aspects of this, but did not provide further details. It can be expected, however, that this will increase the tool's false positive rate. Additionally, it is important to note that the tool is not designed to detect child sexual abuse material that is not part of the CSAM database, i.e. new material being produced or existing CSAM that has not (yet) been detected by law enforcement or clearing houses, such as IWF. Although this limitation is clearly reported by Cyacomb, it will likely lead to a high rate of false negatives when the tool is applied to new or previously unknown child sexual abuse material in a real-life E2EE environment. With regard to the **state-of-the-art**, **robustness** and **scalability** criteria, insufficient information was available for a full assessment. No details about the technology used were shared. What was reported, is that, as the Contraband Filter database is stored remotely, it will not work offline, nor will it work if a user blocks the network connection to the server where matching takes place. With few details provided on the implementation, we can not confidently say it will perform with a poor network connection. Furthermore, insufficient information was made available to assess the tool's performance when applied in different E2EE environments. Likewise, no description of the data distribution in the CSAM database used by the Contraband Filter was made available to enable the assessment of the **fairness/non-bias** criterion. We strongly advise a collaboration with IWF (or other organisations that have access to the aforementioned CSAM database) to examine any bias in the database, describe its limitations and develop bias mitigation measures during the further development of the PoC tool.

From a **security** perspective, Cyacomb mentions an independent security audit of their Contraband Filter technology from specialists Advent IM and First Response regarding the robustness of their filter to "look-up" attacks (in which hashes are used to "look up" content using a search engine or peer-to-peer filesharing system to retrieve the original data), but the evaluation team did not have access to the results of this audit, nor the details of the system itself. Regarding the risk of bad actors testing their content and deriving information that could enable them to circumvent detection, this is reported to be mitigated by collecting metadata on client behaviour (which is in turn protected by Privacy Assured Matching). However, Cyacomb did not elaborate on any measures against potential poisoning attacks to the Contraband Filter.

Moreover, a key security issue identified by the team is the PoC tool's ability to scan everything that is in the purview of the E2EE application. For example, scanning happens oblivious to the user. This means that access to user memory regions happens without the explicit knowledge of the user — the project reports mention that any result of scanning can be hidden from the user. Such technology can be abused by re-purposing it to have unhindered access to user devices with system-wide privileges. Finally, should Cyacomb use perceptual hashing in a future development of the project, it should consider potential attacks to its perceptual hashing techniques (see e.g., [6, 9]).

Regarding **disputability** and **accountability**, no measures to mitigate potentially unfair outcomes can be discerned from the available documentation (e.g., human review before reporting, or appeal and/or redress mechanisms in case of false positives). Cyacomb notes that they researched opportunities to improve secure privacy reporting to an authorised reporting authority based on more efficient approaches to discarding irrelevant and false positive reports, including approaches to preventing malicious activity such as false reporting to the reporting authority. However, we did not have access to the outcome of this research, nor do we know the extent to which they were implemented in the PoC tool.

Finally, no measures to inform people that their communications are being screened, blocked and potentially reported are contemplated. Whilst this, in and of itself, is not a problem insofar as the operator of the E2EE service implementing the PoC tool must provide this information in its capacity as data controller, it would be ideal that PoC tools are embedded with **transparency** measures intended to make abundantly clear that the operation of a screening/blocking tool in an E2EE environment is taking place[27].

5.2 Project SafeToWatch

SafeToWatch is supplied by SafeToNet and the Anglia Ruskin University. The SafeToWatch PoC tool is an on-device moderation technology that performs machine learning based content analysis while the camera application is being used, to prevent the creation (i.e. photographing and filming) and later dissemination of nudity, violence and pornography. The machine learning inference runs locally on a device without the need for cloud interaction. Within the framework of the project, SafeToNet aimed to further develop their on-device technology so that it could be trained to identify CSAM-related content and monitor and protect incoming content consumed by common mobile device apps.

From a **human-centred** perspective, this PoC tool takes an approach that directly tackles one of the drivers of the online CSAM issue whilst providing users with an adequate degree of agency as to whether they want to be subject to the scanning and blocking of their images and videos. SafeToNet rightly observes that there is a lot of material that is innocently created by children or teens which can be damaging when it falls into the wrong hands. Therefore, they sought to prevent the creation of CSAM-related content at source: an SDK is installed on device and integrated with the camera app, and machine learning models will assess what appears on screen to prevent the creation of abusive material. Since the analysis will be performed on device, the confidentiality of the user's content is preserved. Moreover, there is no reporting mechanism contemplated (only blocking), which reduces the likelihood of an attacker intercepting private communications. Most importantly, user (child or parent) consent is required to scan images and videos on device. In particular, users are clearly advised about what the PoC tool is doing, which gives them the possibility to refuse its use. In this way, no user is obliged to be subject to surveillance or censorship against their will, which is consistent with the values of autonomy, self-development and agency that human rights are called upon to promote. In its current status, once adult sexual content is detected, a user interface provides users with educational messages about the risks of sharing the content they are trying to send. This approach

[27] This is in line with Article 5(1)(a) UK GDPR and the ICO's guidelines on transparency: "Transparent processing is about being clear, open and honest with people from the start about who you are, and how and why you use their personal data." See Principle (a): Lawfulness, fairness and transparency, at `https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/lawfulness-fairness-and-transparency/#transparency`

could contribute to guiding the behaviour of children and teens who are beginning to explore their sexuality, as they are nudged into reflecting on the potential harmful consequences of their actions in this regard. However, this would require further research to be confirmed. Under this configuration, the interests of all individuals involved are being considered: users keep enjoying the confidentiality of their content and have agency as to whether or not to be subject to the PoC tool's restrictions. Perceived offenders, however, would only see their content blocked after they give permission to the PoC tool to work on their devices, which is unlikely to happen.

Moreover, there could be additional caveats. Firstly, SafeToNet explains that it aims for its technology to be deployed by apps like video conferencing platforms, gaming platforms, and even workplace managers and network providers. In these scenarios, there is a huge risk that user consent to get one device's camera monitored and photo library scanned be buried within the relevant app's terms and conditions, in which case the degree of agency provided by the original user consent requirement would be lost. In other words, if an app imposes upon users the SafeToWatch software as a precondition for its use (as is customary in online settings), the requirement of user consent becomes an illusion of choice and control. Crucially, if SafeToWatch were to be widely adopted by digital services in the aforementioned way, the automated scanning and blocking of content would be *de facto* imposed across the board.

Secondly, SafeToWatch also explains that they intend to develop their technology further with reporting in mind, so that, for example, video conferencing platforms can report live-streamed CSAM by using SafeToWatch in their technology stack. In this scenario, safeguards such as human review prior to reporting and redress mechanisms for cases of wrongful detection of CSAM should be embedded in SafeToWatch, not least in consideration of the risk mentioned in the preceding paragraph.

Based on the available documentation, it is not possible to give a final answer as to whether SafeToWatch fulfils **human rights** law's necessity and proportionality requirements to justify its potential impact on people's fundamental rights and freedoms, particularly the rights to privacy and freedom of expression. This is because the extent to which the PoC tool is effective, and therefore necessary, is not (yet) clear. Nevertheless, the proportionality standard is to some degree met on account of certain positive features in the PoC tool's design.

To be necessary, a restrictive measure must be effective to attain its stated objective. It follows that the PoC tool's models must be capable of distinguishing between what is CSAM and what is not. At the point of evaluation SafeToWatch models were yet to be trained with real CSAM data, so adult sexual content (on which the tool is currently trained) is used as a proxy for CSAM for the purpose of this assessment. As mentioned in the Human Rights Impact sections of the preceding PoC Tools, machine learning based filtering and detection tools tend to be prone to errors, and in the absence of any performance rates reported for this PoC tool, it can only be assumed that its models are no exception. Human review is essential to keep errors to the minimum possible, yet no human intervention is contemplated in this PoC tool's operation. Also, the PoC tool must be the least intrusive measure amongst equally effective measures. Given SafeToWatch's innovative approach focused on objectionable content production — as opposed to merely distribution — and user education, it is difficult to find a benchmark of equally effective technologies. Arguably, client-side scanning and blocking of known CSAM could be one of these measures, although this is debatable.

With regard to proportionality, the scope of this PoC Tool's impact is not excessive, which is one of its most positive traits. The PoC Tool does not affect every user of an E2EE service — or every person having a mobile device — but only those who willingly consent to have their device cameras intervened and images and videos scanned. There is a big difference between *imposing* the use of a monitoring technology and *giving the option* to use it. This PoC tool's limited scope means that only users who want to protect themselves (or their children) will consent to see their fundamental rights and freedoms encroached upon. This is in stark contrast to other solutions which compromise the privacy of every user of the relevant service — including those within vulnerable groups, journalists, activists, and civil society actors — based on a universal

(and unfounded) suspicion of CSAM-related engagement. In turn, the extent of the PoC tool is admittedly broad, as it involves the scanning of all images and videos to be created and currently stored on device — which are bound to include sensitive data revealing sexual preferences. However, this is not the broadest extent we have seen: private messages and texts in E2EE services are not affected at all.

On the data protection front, things are somewhat unclear. On the one hand, the PoC tool asks for consent prior to installation, and informs users on what the technology does, the data it accesses, how it is handled and where it is stored. Therefore, it relies on a legal basis for processing that promotes individual control and informational self-determination, which is a highly welcome development. In addition, this disclosure of information is in line with the **transparency** principle and individuals' right to obtain information about the processing. However, the right of individuals not to be subject to a decision based on a solely automated system means that there should be human intervention if a user so requires (see Article 22(3) UK GDPR), but the reports do not contemplate that intervention in any stage.

In turn, under the current configuration, the PoC tool does not capture or transfer images or videos of any kind, not even to a reporting server. This architecture greatly protects the confidentiality of users' personal data and thus reduces significantly the likelihood of re-identification. However, the PoC tool is reported to use a combination of supervised, semi-supervised and unsupervised machine learning techniques (no further details were disclosed). It is unclear, based on the available documentation, whether or not the images and videos of those who gave access permissions to the PoC tool are used to improve the tool's models. In the affirmative, use of personal data in this way should be made explicit and clear at the time permissions are requested.

There are two final observations. If SafeToNet moves forward with its plan to implement reporting of content to authorities, any transfer of data must be in strict compliance with the UK GDPR's data quality principles and associated requirements, and this functionality must be clearly explained at the time user consent is requested. In addition, not least on account of freedom of expression considerations, safeguards such as human review, an appeal process, and redress mechanisms (cf. **disputability** and **Accountability**) for cases of erroneous detection and reporting of CSAM should be implemented.

With regard to the **security** criterion, the PoC tool runs locally on-device, relying on inherent device level security systems to impede malicious users. Additionally, SafeToWatch is being developed so that it can be deployed by other technology companies within their own devices and applications. Hence, the project also relies on security and penetration protection measures employed by those companies. Nonetheless, SafeToNet's main goal is to detect CSAM using machine learning. Recent literature discussed a range of security considerations vis-à-vis the potential vulnerability of trained models to adversarial machine learning (e.g., [19]. For example, a malevolent actor could attempt to manipulate the training of a machine learning model to intentionally misclassify any input with an added trigger (i.e. backdoor attacks [21]), or to poison the model aiming to make it misbehave on specific inputs (e.g., [26]). To the best of our knowledge, these aspects are not discussed in SafeToWatch's reports. We highly recommend considering such security concerns when training and testing on actual CSAM is able to commence.

At the point of evaluation, SafeToNet had not been able to train or test their PoC tool on actual CSAM data. As a result, we were not able to evaluate the **performance**, **robustness**, **scalability**, **state-of-the-art** and **fairness/non-bias** criteria for this PoC tool. The reported accuracy for the prototype of SafeToWatch for detecting adult explicit content (not CSAM) was 98.2%. This was achieved on a sample size of 6,000 adult sexual content files. Without further information on the experimental details, it is impossible to adequately assess this reported performance. We strongly advise incorporation of measures ensuring fairness/non-bias when access to actual CSAM-related data is achieved and to report on any dataset and model limitations in a transparent way.

5.3 Project GalaxKey

Project Galaxkey is supplied by GalaxKey and Image Analyser, Yoti. The proposed solution aims to combine three commercially available products: the Galaxkey encryption platform for providing end-to-end messaging, Image Analyser for performing AI-based explicit content filtering and Yoti for providing age estimation and digital identity verification. The intended PoC tool is described as a new E2EE platform that would require users to register using Yoti to determine their age, scan any text message for profanity, analyse the content of each attachment for explicit content, and send it securely to other Yoti verified users. Scanning is done pre-encryption, allowing for the IWF API to be included optionally.

In its efforts to balance the democratic need of having private and secure communications with the need to prevent the misuse of E2EE technologies for child sexual abuse-related purposes, Galaxkey's PoC tool is excessively tilted to the latter need. On the one hand, the PoC tool's on-device moderation approach ensures that the private communications of both law-abiding users and CSAM victims remain confidential and inaccessible to potential adversaries. Further, when the machine learning systems detect an explicit message, the perceived offender's email, IP address and relevant content is sent in encrypted form to a reporting server (Galaxkey Secure Workspace), where it is claimed data is kept securely, only authorised personnel have access to it, and access to that data can be audited and tracked. In this way, the confidentiality of perceived offenders' communications is to some extent preserved, and the risk of re-victimisation due to CSAM leakage and dissemination is reduced. Moreover, the PoC tool's design significantly mitigates the risk of malicious users implicating others. In particular, there is a requirement of digital identity verification, and to mitigate errors in age estimation, users have to identify themselves with a government approved document like a passport or driving license, before using the E2EE platform. As a result, unless a device is stolen and misused, the identity of senders could potentially be identified with a high degree of accuracy.

On the other hand, however, based on the available documentation, there are no safeguards against errors or unfair outcomes other than human moderation at the reporting server. This means that protected speech will be censored at the discretion of a person likely without adequate training on freedom of speech considerations, and in the absence of notification mechanisms and appeal processes, many users will see their content blocked and reported, without any recourse to challenge these decisions. In addition, users of the E2EE service will be unlikely to know that permanent scanning, filtering, blocking and reporting of their private communications is taking place, and what the consequences for the protection of their personal data may be. Although Galaxkey asserts that its PoC tool obtains user consent in accordance with the ICO's guidelines, we have not seen how this is done in practice. At any rate, the pitfalls of consent are well understood [20], so additional **transparency** measures should be in place to ensure that the E2EE service's users understand the implications of using the service. By way of example, when a new image is detected as explicit and illegal, it is uploaded to the Galaxkey Secure Workspace and then added to the detection model for future detection. If the moderator thinks the image is not illegal, then the image is moved to the 'allowed image workspace', and uploaded as a clean image for training the detection model. The question that follows is, based on the PoC tool's consent notice, can users duly understand that their messages and content are being collected for the purpose of training machine learning models, and the fact that once the content enters a model, it is virtually impossible to retrieve? The lack of adequate redress and transparency mechanisms coupled with the PoC tool's degree of potential interference with people's fundamental rights and freedoms (see below) means that the **human-centred**, **disputability** and **accountability** requirements are not met.

Galexkey's PoC tool fails to meet the necessity and proportionality requirements under applicable **human rights** law to justify its potential impact on people's fundamental rights and freedoms. Relying on AI, this PoC tool automatically and constantly assesses, classifies and reports text messages, images and videos of all the relevant E2EE service's users in order to detect and block profanity and CSAM. Therefore, the scope and extent of this PoC tool's impact is as wide as it could possibly be. Every piece of content intended to be sent privately by every user of the E2EE service is monitored, in such a way that everyone is treated as

a suspect of CSAM-related law violations. Moreover, special categories of data (e.g. religious views, ethnicity, sexual preferences, health conditions) are bound to be contained in users' private communications, yet every piece of content is equally processed, sorted and used to train machine learning models, without distinctions, let alone additional safeguards for sensitive data.

To determine whether a measure is necessary, its effectiveness for achieving its intended goal must be established. This means that the PoC tool's models must be capable of identifying the nuances that distinguish protected from unprotected speech, as well as unlawful from lawful content. However, the COVID-19 pandemic showed that the AI-supported tools of technology giants are not yet up to this task (see [23]), and since this PoC tool's performance on actual CSAM-related data had not been tested at the end of the project (see below), it is fair to assume its models are prone to some degree of error which, in the absence of adequate safeguards, would not be suitable from a human rights perspective. Moreover, since this PoC tool involves the systematic monitoring and analysis of everything contained in private communications, it raises concerns regarding mass surveillance if deployed at scale, and consequently it cannot be deemed the least intrusive measure to attain its goal. Most importantly, given that widespread deployment of this PoC tool would effectively compromise everybody's privacy — including that of those within vulnerable groups, journalists, activists, and civil society actors — technical, legal, operational, and/or contractual safeguards to impede the re-purposing of this technology should be in place. None can be found in the available documentation. This is especially pertinent as the PoC tool's AI models can reportedly detect pornography, extremism, graphic violence, drugs, alcohol, weapons, gambling and risqué material. Hence, the re-purposing of this PoC tool is likely seamless, which is concerning, in the light of regulatory developments both in the UK and overseas[28].

With regard to data protection, it is claimed that users' data is stored in encrypted form in the Galaxkey Secure Workspace, access to this data is controlled, auditable and traceable, and security measures to prevent employees from downloading the data are in place. In addition, user's data is encrypted before being sent to the reporting server. In the light of this, this PoC tool is generally in line with the integrity and confidentiality principle and data security obligations under applicable law. However, to the extent that the entirety of users' communications (i.e. messages, images and videos) is used to update the PoC tool's machine learning models, and an undefined amount of metadata is also sent to the reporting server to derive more information about perceived perpetrators, the PoC tool is not in line with the data minimisation principle. Moreover, it is unclear whether the purpose limitation and storage limitation principles are complied with. There are no assurances that users' data will not be used for other purposes (e.g. to train pornography or violence detection models), nor is it stated for how long users' personal data will be kept. Crucially, since on account of the nature, scope, context and purpose of the processing, a high risk to the E2EE service users' fundamental rights and freedoms is bound to arise, and thus a data protection impact assessment should have been conducted.

Lastly, the impact of this PoC tool on the right to freedom of expression deserves close attention. Since this PoC monitors and analyses every piece of users' private communications before allowing them to be sent, every message is treated as a potential law violation. Prior restraints thus become the norm. And even under the assumption that this PoC tool's models are as accurate as those implemented by large technology companies in their flagship services, there is a huge risk that some users' messages, texts or videos be misclassified and therefore blocked. To be to some extent admissible, censorship of this kind and magnitude should be supported, at the very least, by notifications and appeal processes via which users could challenge censorship decisions they deem unjustified. The only safeguard we were able to identify in this connection was the intervention of a human moderator at the reporting server, who enjoys absolute discretion to deter-

[28] For example, this PoC tool could be repurposed to comply with 'proactive technology' requirements imposed by Ofcom under Section 116 of the UK Online Safety Bill, which applies to illegal content, children's safety and fraudulent advertising. Similarly, repurposing could take place to give compliance to Article 5(2) of Regulation (EU) 2021/784 on addressing the dissemination of terrorist content online, which imposes on hosting service providers the obligation to take 'specific measures' to protect their services against the dissemination of terrorist content to the public.

mine what amounts to protected or unprotected speech, and what is the defining line between lawful and unlawful content. However, even assuming that moderators have been adequately trained on these human rights considerations — something which is yet to be established — this safeguard falls short of the requisite proportionality standard.

Galaxkey is aware of the tool's lack of compliance with ICO and reported seeking advice on the necessary compliance measures, including in relation to the Privacy and Electronic Communications Regulation, anonymisation standards, and how best to structure consent processes.

With regard to the **performance**, **robustness**, **scalability**, and **fairness/non-bias** criteria, insufficient information was provided to perform a useful evaluation. The project proposal mentions the use of various metrics (e.g. Kappa, Cronbach's Alpha, Level and Category Distribution) to measure inter-annotator agreement when labelling data and accuracy for evaluating the performance of their CSAM detection system, aiming for a 70% accuracy. However, no results or details about the training/testing datasets, or other vital experimental details, were shared with the team. We strongly recommend them to report on any dataset and model limitations in a transparent way during the further development of their tool.

Finally, there is insufficient detail provided to evaluate if the PoC tool meets the **security** criterion. Adversarial machine learning attacks as mentioned in the previous section (e.g., a user attacking the model offline to find weaknesses) appear to be possible. We highly recommend considering such security concerns during the future stages of the tool's development. Additionally, since the project proposes an E2EE application with built-in scanning features, threats arising due to lack of end-point security can be expected. There is an expectation that service users can generate and store their own keys. However, no details on key generation, management and revocation were made available.

5.4 Project DragonflAI

Project DragonflAI is supplied by DragonflAI, Yoti. DragonFlAI's PoC tool involves on-device moderation within an E2EE system by combining a machine learning model that detects nudity with an age estimation model. Pictures containing both nudity and an underage person cannot be sent, and the user trying to do so can be flagged. Thus, this PoC tool does not rely on databases of known CSAM. Although the aim of this PoC Tool is to dramatically reduce the need for human moderation, any potential issues — i.e. errors — can be flagged and sent through for human moderation if users feel content is incorrectly flagged and the app provider using this PoC contemplates this option.

The on-device moderation approach of DragonflAI's PoC tool seeks to balance the societal need to have secure and private communications with the need to fight their misuse for child sexual abuse purposes, with a fair degree of consideration for the interests of all individuals concerned. However, its scope is broad and there is a clear risk of utilisation for widespread surveillance and censorship of protected speech.

On the one hand, it requires no client-server communication to detect and block CSAM, and therefore the private communications of E2EE services' users are protected against potential adversaries snooping on the network. Also, this PoC tool is intended to be fully autonomous, i.e. it can moderate content without relying on a third party, both server and human. This trait strongly mitigates the risk of CSAM leakage — at least before the relevant content reaches the competent authorities — and as a result the likelihood of re-victimisation is reduced. The autonomy of this PoC tool also means that no humans are exposed to the disturbing content of CSAM, thus contributing to the overall level of people's psychological well-being.

On the other hand, this PoC tool uses the combination of nudity and the presence of an underage person as proxy for CSAM. In particular, the tool detects nudity and estimates the ages of all faces featured in an image. Images where a child's face along with nudity is present are flagged, but CSAM cannot be detected in media in which children's faces are not visible. Aside from the fact that machine learning models intended to find new CSAM, like other machine learning approaches, are likely to be prone to errors, and the PoC tool's

accuracy rates in fact are bound to cause a fair amount of false positives if deployed at scale (97.9% accuracy at nudity detection and around 1.5 years of mean absolute error in age estimation), the presence of nudity and an underage in an image is a poor proxy for CSAM (a mother taking a picture of her newborn son having a bath can be wrongly classified as CSAM). Also, there seem to be no safeguards or measures to verify the correctness of the PoC tool's results, to ensure transparency of the fact that scanning and filtering is taking place, or to redress unfair outcomes. The absence of human moderators means that false positives may be reported. An error of this type can have severe consequences for the sender, who could be reported as potentially having committed a very serious crime and have her private content and personal data processed without knowing. Whereas it is ultimately up to the app provider to determine whether or not to use human moderators and/or automatically report content, the likelihood of occurrence of the aforementioned scenario in real-life coupled with the lack of minimum safeguards to prevent and redress it is inconsistent with the **human-centred** approach a CSAM-detection tool must have.

The impact of DragonflAI's PoC tool on people's **fundamental rights and freedoms** is neither necessary nor proportionate to the intended aim of protecting children against sexual abuse. In particular, the tool involves the automated filtering and scanning of all the relevant E2EE service's users' content data, with a view to detecting CSAM images that have not been previously found. Thus, the scope and extent of this PoC tool is wide. Instead of targeting people under investigation based on specific, reasonable and individual-level suspicion, the privacy of all of the relevant service's users' private communications is compromised, and everybody is effectively treated as a suspect of looking at or spreading CSAM.

In addition, since the automated analysis of users' content is performed to find previously uncovered CSAM based on nudity plus age as proxy for it, this PoC tool can be deemed neither effective nor the least intrusive measure to achieve its purpose of preventing CSAM dissemination. Instead of detecting and reporting images in respect of which there is a high degree of certainty as to their unlawful nature (as is the case of confirmed instances of CSAM), all users' images are processed, classified, labelled, and potentially reported by a fully automated decision-making system which is not exempt from error. Also weighing against the proportionality of this PoC tool is the fact that fundamental rights and freedoms other than privacy are at stake, as the continuous scanning and evaluation of all users' images poses a direct threat to individuals' freedom of expression. There is a great chance that an individual's speech can be mistakenly classified as unprotected, thus potentially exposing such individual to scrutiny on the part of law enforcement agencies if human moderation is not enabled. Over time, this risk of exposure and scrutiny is likely to have a chilling effect on lawful speech. Crucially, the absence of transparency safeguards means that users will be unaware of the PoC tool's operation, which is in contradiction with the principle of legality that individuals must be able to know what restrictions are applied to their protected speech.

Finally, with regard to data protection considerations, data protection requirements are intended to be managed by the service that employs the PoC tool. Whilst this may be an acceptable solution in terms of compliance — as the PoC tool's operator can be deemed a data processor — it scores poorly in terms of the actual impact on the right to data protection that use of this PoC tool is likely to have. DragonflAI recommends that an image be only uploaded or sent to other users after the PoC tool has analysed it, and if it is found to be illegal, the company using the PoC tool has the choice to send the image directly to authorities, or to simply deny upload before the image leaves the device. This design offers no assurances whatsoever that individuals' personal data (i.e. the images found to be illegal) will be shared with authorities based on adequate security measures, that it will be used only for the purposes of CSAM detection, or that it will be stored only for as long as strictly necessary for such purpose. Nor are there any assurances that individuals will be able to assert their data protection rights effectively, especially their right to human review for decisions made solely on the basis of automated processing. On the plus side, this PoC tool complies with the data minimisation principle, as the only data processed by it are images.

Finally, the limited availability of safeguards against errors or unfair outcomes *by design* (cf. **disputability** and **accountability**) caused by the PoC tool confirms its potential disproportionate impact on individuals'

human rights. Given that the reporting of CSAM after automated detection may significantly affect the data subject concerned, no report should be based solely on the outcome of automated detection, especially in consideration of the rather high likelihood of errors arising from the poor proxy for CSAM that was chosen. However, one of the selling points of this PoC tool is its potential autonomous operation, i.e. without human involvement. As a result, if it were to be implemented as such by the app provider, no human would be confirming the accuracy of the PoC tool's decisions, and thus users whose images have been erroneously blocked or reported would have no recourse to challenge this outcome in the latter scenario.

With regard to the **performance**, **robustness**, **scalability**, and **fairness/non-bias** criteria, we were not able to perform the evaluation due to the lack of information provided. Based on the project description, the DragonflAI PoC tool intends to use two independent machine learning based systems, each of which have limitations. We expect that a key limitation will be caused by the age estimation process, which requires the extraction of a face from the image, which if missing from the image, will lead to a false negative. To circumvent detection, CSAM creators can easily avoid or obfuscate their victims' faces. Finally, a limitation arising from the combination of both systems for CSAM detection or prevention is the potential blocking and reporting of legal images and videos that depict (partially) undressed children, such as described in the bath time example above. These aspects are not discussed in the reports and could be mitigated by training and testing on actual CSAM, rather than the combination of nudity and the presence of an underage as proxy.

Given the PoC tool runs on the device and must process everything that appears on the screen, in (near) real time for it to be effective, a clear trade-off in performance and resources is to be expected. A highly accurate model that can run in (near) real time on low-end devices, while being mindful of resources, seems extremely ambitious. For example, constant monitoring (and thus by inference using the machine learning models) of the device will have significant power consumption needs. Attempts to reduce this will likely impact the tool's performance. Additionally, it can be expected that when deployed to a significant number of users, the PoC tool's expected false positive rate would cause the reporting mechanisms to become overloaded given the volume.

With regard to **security**, the machine learning models used by the PoC tool, again, could be vulnerable to adversarial Machine Learning attacks (see previous sections), but no security measures are discussed within the reports. Moreover, DragonFlAI is part of a host E2EE application and performs matching in the device itself. The application is further complemented with age verification from Yoti. Such an arrangement entails a complex update system in response to inevitable bugs. The proposal does not discuss the patch management (in response to inevitable bugs) from each of the individual vendors.

5.5 Project T3K

Project T3K is supplied by T3K-Forensics. T3K Forensics' PoC tool aims to detect CSAM on device with AI-based classifiers, which are trained based on picture and video content. It follows a two-layer approach, the first is the detection of pornography/nudity, and the second is determination of the presence of children in the screened content. Machine learning classifiers also estimate facial age and gender.

In balancing the societal need to have secure and private communications with the need to fight their misuse for child sexual abuse purposes, T3K Forensics PoC Tool placed excessive emphasis on surveillance, disregarding the consequences of this practice whilst implemented at scale. Nonetheless, there are both positive and negative aspects to highlight.

On the one hand, it operates on device, so it requires no client-server communication to detect and report CSAM. Consequently, the private communications of E2EE services' users are protected against potential adversaries snooping on the network. Also, this PoC tool allows for the setting of a threshold of the amount of detected content, to eliminate the threat of a person sending just one file to another person to implicate them. In this way, reports of false positives are reduced. In addition, whilst this PoC tool uses the

combination of nudity and the presence of an underage as proxy for CSAM — which, as noted above, is a poor proxy for CSAM — reported CSAM is intended to be manually verified before the E2EE service takes action. This verification removes the risk that people sending legal pictures of their children be reported to the authorities on grounds of being suspected of committing a crime.

On the other hand, machine learning models trained to detect new CSAM are likely to be prone to errors, and the PoC tool's performance has not been reported. As a result, it is fair to assume that this PoC tool might generate a significant amount of false positives. In addition, users of the E2EE service will be unlikely to know that permanent scanning, filtering, and reporting of their private communications is taking place, and to what extent their personal data may be compromised. Based on the available documentation, it is not possible to determine whether users' images and videos will be used to update and refine the PoC tool's machine learning models. Therefore, **transparency** measures should be in place to ensure that the users of the relevant E2EE implementing the PoC tool understand the consequences of its use. Also, the available documentation is not clear as to the extent to which end-users can control the risks they are exposed to. For instance, can they set app permissions, and with what granularity? Can end-users opt out of scanning? These may be questions that concern the user-facing application that would be embedded in this tool. Nevertheless, a complete evaluation of the **human-centred** criterion would need a clearer picture of user control.

The impact of T3K Forensics' PoC tool on people's **fundamental rights and freedoms** is neither necessary nor proportionate to the goal of protecting children against sexual abuse. More specifically, to detect and report CSAM, T3K Forensics' PoC tool filters and scans through automated means all the relevant E2EE service users' content data. Consequently, the scope and extent of the this PoC tool is very wide. The PoC Tool does not target people based on specific, reasonable and individual-level suspicion. Instead, it compromises the privacy of all of the relevant service's users' private communications, treating every user as suspected of looking at, or sharing CSAM.

Also, as the automated analysis of users' content is performed to find new CSAM based on nudity plus age as proxy for it, this PoC tool cannot be effective to attain its purpose of preventing CSAM dissemination. All users' images and videos are processed, classified, labelled, and potentially reported by an automated decision-making system with a potentially significant likelihood of error. While human moderation can ensure that false positives are not reported to the authorities, it is questionable whether moderators will always verify false positives when the relevant E2EE service has a large user base and moderating content at scale becomes unmanageable (cf. **disputability** and **accountability**). Crucially, if this PoC tool is deployed at scale, the systematic monitoring and analysis of users' content would amount to mass surveillance. Consequently, it can be hardly be deemed the least intrusive measure to attain its goal. Widespread surveillance means that everybody's privacy — including that of those within vulnerable groups, journalists, activists, and civil society actors — would be impacted upon. Therefore, technical, legal, operational, and/or contractual safeguards to prevent the re-purposing of this technology should be explicitly contemplated. However, none can be found in the available documentation. T3K Forensics has a general Object Detection model which can reportedly detect guns, terrorist symbolism and other specific content. Accordingly, there is nothing preventing this PoC tool be re-purposed in the future to detect content other than CSAM, and in fact doing so is likely seamless and inexpensive. The negative impact this would have on freedom of speech is not to be underestimated.

On the data protection front, data protection requirements are supposed to be managed by the service that employs the PoC tool. This design is acceptable in terms of legal compliance — as the PoC tool's operator can be deemed a data processor — however, the actual impact that use of this PoC tool is likely to have on the right to data protection is questionable. Upon detection of CSAM, T3K Forensics notifies the E2EE service provider that there was a hit, and based on that notification the provider can send the content to the authorities. Thus, this design offers no assurance that an individuals' personal data — that is, the images and videos found to be illegal — will be shared with authorities based on adequate security mea-

sures, that it will be used only for the purposes of CSAM detection, or that it will be stored only for as long as strictly necessary for such purpose. Moreover, special categories of data (e.g. religious views, ethnicity, sexual preferences, health conditions) will inevitably be contained in users' private communications, yet every piece of content is equally processed, labelled and potentially used to train machine learning models, without additional safeguards for this sensitive data. Finally, the classifier is effectively a black box which would take unencrypted images and videos to determine the presence/absence of CSAM. Documentation should provide adequate clarity on the judicial oversight involved in the process. Judicial oversight can be translated in systems through effective controls that would prevent bulk surveillance. From the perspective of the victims, the following sentence needs more clarification: "If a person only receives very few files, only once, this will not be reported."

The solution relies on the **security** of mobile operating systems to prevent data leakage. With respect to apps stealing from other apps, the proposal should make a distinction between the protections in iOS and Android while outlining their protection mechanisms.

Furthermore, T3K Forensics did not provide any explanation on how they intend to limit the purpose of the scanning technology to specific regions of memory and who decides that. This also has implications for human rights (see above). To that end the privileges that the scanning component will assume once deployed needs to be explained.

So far as adversarial tendencies to evade machine learning attacks are concerned, we recommend that T3K Forensics elaborate on their plans to make their classifiers resilient against minor perturbations (see e.g., [**?**]). They proposed the use of metamorphic classifiers to reduce adversarial attacks by users. This works by giving different users different versions of the same classifier. Although this increases the difficultly for a set of attacks, dedicated attackers may still be able to bypass this as each model is stored locally on the device. For example, metamorphic classification is unlikely to provide any security against sybils.

With regard to **performance**, T3K Forensics discussed the concept of false positives, recall and precision, but did not report specific values, nor experimental details. False positives are defined as cases where an image or video contains a scene where the system incorrectly classifies it as containing a naked body (in a pose 'relevant' for CSAM) with an age estimated to be 'pre-pubescent' (and/or under 18, it is unclear), on 'biological features of pre-pubescent persons'. Although the project proposal mentions that their underlying CSAM detection technology is already available and in use[29] and thus part of an end product, details about the performance, the system's limitations and a description of the datasets used to train and test their models were not made available.

Finally, as with some of the other approaches, there is a clear trade-off to be made between the performance rate and the processing time and resources. T3K Forensics proposed a queuing mechanism that delays scanning until the device is connected to power as a means of reducing the impact on the user. Nonetheless, the deep neural network based approach must still function on low-end devices and thus it is expected to suffer a drop in performance. Given the system runs on the device, the PoC tool can be expected to work offline or on a poor network condition.

6 Conclusion

Along with a description of this framework, that was developed with feedback from members of the cyber security & privacy community along with stakeholders from academia, industry, law enforcement and NGOs focusing on online child protection, in this report we presented a case study in which we performed a qualitative analysis of five Proof-of-Concept tools against this framework. More specifically, we described the presence or absence of different measures that assure compliance with our evaluation criteria.

Given (1) the exploratory nature of each PoC tool and (2) the publication of the evaluation criteria post the final delivery date of the projects, it was to be expected that not all criteria could be met in full. Hence,

[29] We assume it is part of their Law Enforcement Analytics Platform (LEAP).

the evaluation presented in this work is intended as a guide for the safety tech industry to positively influence the development of automated tools for online child protection, while also ensuring all users benefit from such solutions.

Nonetheless, the present evaluation does provide some noteworthy insights into the difficulties of building online child protection tools, especially in the highly challenging context of E2EE environments. First, 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 for most of the PoC tools. Although none of the PoC tools proposes to weaken or break the end-to-end encryption protocol, the confidentiality of the E2EE service users' communications cannot be guaranteed when *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. This differs significantly from automated CSAM detection tools that are currently being used by law enforcement agencies in their investigative practice pertaining to online child protection.

Secondly, although advertising the potential re-purposing of a tool or machine learning model seems valuable from a commercial point of view, it is highly concerning in the context of analysing protected communications. Therefore, it is essential to include technical, legal, operational, and/or contractual safeguards to prevent the re-purposing of such technologies prior to any deployment in a real-life E2EE application.

Third, transparency, disputability and accountability proved to be problematic in most of the PoC tools. Additionally, none reported any maintenance strategies (aside from collecting data for retraining/finetuning the tools). Despite not being end products, these principles should be taken into account by design, rather than relying on the scrutiny of the platforms in which they might be integrated.

Finally, the key limitation of this evaluation has been the absence of detailed experimental information due to confidentiality issues. As a result, none of the PoC tools could be assessed for their fairness/non-bias, performance, use of state-of-the-art techniques, robustness or scalability. Hence, our future work will include examining how the evaluation framework presented in this report can be further amended and refined to enable the establishment of ethically responsible benchmark datasets for developing and evaluating online child protection tools. This way, such technologies can be evaluated independently without the risk of compromising commercial interests.

7 Discussion

The Safety Tech Challenge was directed at the development of technical prototypes, but the context of application, and the potential for subversion and false positives and negatives, remains an important factor to consider if technical solutions are to be effective.

Tariq et al. [24] reviewed tool development from a 'human lens' perspective. They critique evaluation measures that only report on the success of the selected algorithmic approach, using measures such as precision, accuracy and recall. A key finding from their analysis of 45 papers published between 2008–2018 is that none report user evaluation by the child. The primary model applied to the detection across the decade was nudity/skin detection, with 86.4% of papers deploying computer vision techniques and only 9% using natural language processing (NLP). Two of the 45 papers focused on mobile device detection, with the remainder producing general solutions. A significant conclusion is that all the approaches were applied post-production, detecting risk, rather than risk mitigation that prevents production.

Three trends in technical development mitigate against comprehensive preventative efforts. Firstly, the separate trajectories of technical research directed at detecting CSAM (reported in the academic international peer-reviewed literature) and child protection knowledge (largely covered by NGO and policy reports)

and secondly a disconnection between academic and private sector research in this field. Thirdly, end-to-end encrypted (E2EE) environments, dark web fora, along with P2P networks have grown rapidly.

Except for the Bracket Foundation [2], NGO and policy reports focus on the problem of online sexual abuse of children, highlighting its characteristics and prevalence as an industry, an organised crime enterprise, a product of sex tourism, the role of self-generated and social media, the trauma and impact on children, and so forth. These are the contexts in which CSAM is produced with often serious consequences for the children involved. Recommendations generally call for integrated policies and responses, at a national and regional level, in the hope that these will deter perpetrators and assist victims. The technical literature reports on a range of different models to address CSAM detection, including webcrawlers, filename and filepath analyses, chatbots and various age and skin detection techniques, often paying little attention to the application of tools by law enforcement and the real-world challenges of evidence gathering, prosecution and child protection. As accuracy, precision and recall in the detection of CSAM are greatly improved by deep learning models, the massive scale of retrieval presents law enforcement and clearing houses such as NCMEC and IWF with a significant challenge; how to manage the volume of CSAM? The Bracket Foundation [2], Burszstein et al [3], and Sanchez et al [22] recommend an emphasis on improving computational approaches to the law enforcement processing task. These technologies must address the enduring problem that processing must also fit with evidential requirements, which vary across the world. In this context, the recent focus (particularly in Europe following implementation of GDPR) on improving transparency and explainability, presents a well-known dilemma in the field: how to make the tools effective and at the same time do not report sufficient detail that perpetrators find ways to circumvent them, or compromise commercial interests?

Secondly, almost all the tools in use (such as those identified by the Bracket Foundation (ibid, 2019) are commercialised. Whilst they may or may not be effective, evaluations of effectiveness are (a) not provided or (b) lack clarity on the specific purpose within the spectrum of CSAM. If it is the detection of CSAM, this will not effectively prevent the production of CSAM or further victimisation; partly for resource reasons (as noted above) and partly for displacement reasons. Internet service providers and platforms may manage to reduce CSAM on their sites by using computational tools, but the sheer volume reported to law enforcement means investigation cannot keep up. Further, as the authors of [12] observe, even where platforms such as Google disrupt and deter CSAM, perpetrator activity is merely displaced to sites and jurisdictions where no such deterrence is in place.

Thirdly, private traffic such as in E2EE environments, dark web fora, and P2P networks has grown rapidly. Whilst generally benign, serving many important functions that protect privacy and human rights, these environments also provide an unprecedented nurturing and supportive digital space for creating, sharing and disseminating CSAM. To date, there has been no published research on computational tools that can prevent CSAM in E2EE and limited studies on P2P and the dark net. With the exception of bots designed to lure and/or deter CSAM offenders [25], all technical approaches are currently designed to detect CSAM post-production [12]. Thus the PoC tools reported here that provide a context to mitigate the creation of CSAM and prevent its publication prior to uploading are innovative.

A continuing problem, noted in research and recurring in this study, is access to CSAM on which to test tool development, which most researchers and safety tech developers lack. Only one fifth of tools developed over the last five years and reported in peer-reviewed sources managed access through working in partnership with law enforcement or clearing houses [17]. Even where access is granted, evaluation is limited to testing the accuracy of the tool.

Evaluation of effectiveness is challenging given the range of different models and approaches. Where tools are developed specifically for the detection of CSAM, either directly or indirectly through age and nudity detection, accuracy is the most commonly reported assessment criterion. Amongst tools developed since 2016 this ranges from 60% for the detection of young children [1] to 97% for filepath analysis [18]. Re-

call is also reported in some cases ranging from 65% for a stepwise law enforcement processing model [15] to 94% for filepath analysis [18]. Other methods included calculating Mean Average Error (for age estimation and CSAM severity detection), and Goodness of Fit (network analysis) [17]. Performance metrics alone cannot fully evaluate efficacy in CSAM prevention and consideration must also be given to the wider context in which tools will be applied; at what point in the creation of CSAM is the tool directed, who has responsibility for its use and the consequences, intended or unintended, during application?

There are no straightforward answers to these challenges but a greater awareness in the child protection field of computational tool development and technical challenges presented by E2EE, and a reciprocal growth in awareness in the cybercrime/computational tool development field of the practicalities of child protection would undoubtably help. What seems to be lacking, is an agreed international standard for CSAM tool evaluation that is shared by child protection organisations, the private sector, and researchers.

The evaluation framework presented in this report is offered as an initial starting point for this wider enterprise that must involve representatives from across the private, public, and civil society, and include children and young people themselves.

8 Acknowledgements

The REPHRAIN evaluation team would like to thank REPHRAIN's Strategic Board for their time and efforts in supporting our work and the community for providing input during the consultation on the evaluation criteria. We would also like to thank the Safety Tech Challenge Fund for inviting REPHRAIN to undertake the evaluation and the suppliers for engaging with the evaluation.

References

[1] Lacey Best-Rowden, Yovahn Hoole, and Anil Jain. Automatic face recognition of newborns, infants, and toddlers: A longitudinal evaluation. In *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–8. IEEE, 2016.

[2] Bracket Foundation. 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`, 2019.

[3] Elie Bursztein, Einat Clarke, Michelle DeLaune, David M Elifff, Nick Hsu, Lindsey Olson, John Shehan, Madhukar Thakur, Kurt Thomas, and Travis Bright. Rethinking the detection of child sexual abuse imagery on the internet. In *The world wide web conference*, pages 2601–2607, 2019.

[4] DCMS, Home Office, and Public. Safety tech challenge fund supplier guidance, 2021. `https://view.publitas.com/public-1/safety-tech-challenge-fund-supplier-guidance/page/1` [Online; accessed on 22-February-2023].

[5] EU Parliament. Text of the provisional agreement on the digital services act. `https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/IMCO/DV/2022/06-15/DSA_2020_0361COD_EN.pdf`, 2022.

[6] Qingying Hao, Licheng Luo, Steve TK Jan, and Gang Wang. It's not what it looks like: Manipulating perceptual hashing based applications. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 69–85, 2021.

[7] HM Government. The government report on transparency reporting in relation to online harms. `https://assets.publishing.service.gov.uk/government/uploads/system/uploads/`

attachment_data/file/944320/The_Government_Report_on_Transparency_Reporting_in_relation_to_Online_Harms.pdf, 2020.

[8] Interpol. Threats and trends child sexual exploitation and abuse: Covid-19 impact. https://www.interpol.int/content/download/15611/file/COVID19%20-%20Child%20Sexual%20Exploitation%20and%20Abuse%20threats%20and%20trends.pdf, 2020.

[9] Shubham Jain, Ana-Maria Creţu, and Yves-Alexandre de Montjoye. Adversarial detection avoidance attacks: Evaluating the robustness of perceptual hashing-based client-side scanning. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 2317–2334, 2022.

[10] Davinder Kaur, Suleyman Uslu, Kaley J Rittichier, and Arjan Durresi. Trustworthy artificial intelligence: a review. *ACM Computing Surveys (CSUR)*, 55(2):1–38, 2022.

[11] Anunay Kulshrestha and Jonathan R Mayer. Identifying harmful media in end-to-end encrypted communication: Efficient private membership computation. In *USENIX Security Symposium*, pages 893–910, 2021.

[12] Hee-Eun Lee, Tatiana Ermakova, Vasilis Ververis, and Benjamin Fabian. Detecting child sexual abuse material: A comprehensive survey. *Forensic Science International: Digital Investigation*, 34:301022, 2020.

[13] Ian Levy and Crispin Robinson. Thoughts on child safety on commodity platforms. *arXiv preprint arXiv:2207.09506*, 2022.

[14] Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. Trustworthy ai: From principles to practices. *ACM Computing Surveys*, 55(9):1–46, 2023.

[15] Joao Macedo, Filipe Costa, and Jefersson A dos Santos. A benchmark methodology for child pornography detection. In *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 455–462. IEEE, 2018.

[16] Corinne May-Chahal and Kelly Emma. Deepening knowledge of online child sexual victimisation. In *Online Child Sexual Victimisation*, pages 141–158. Policy Press, 2020.

[17] Corinne May-Chahal, Claudia Peersman, Awais Rashid, Maggie Brennan, Emma Mills, Peidong Mei, and John Barbrook. 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, Lancaster University & University of Bristol, forthcoming.

[18] Mayana Pereira, Rahul Dodhia, Hyrum Anderson, and Richard Brown. Metadata-based detection of child sexual abuse material. *arXiv preprint arXiv:2010.02387*, 2020.

[19] Nikolaos Pitropakis, Emmanouil Panaousis, Thanassis Giannetsos, Eleftherios Anastasiadis, and George Loukas. A taxonomy and survey of attacks against machine learning. *Computer Science Review*, 34:100199, 2019.

[20] Neil Richards and Woodrow Hartzog. The pathologies of digital consent. *Wash. UL Rev.*, 96:1461, 2018.

[21] Ahmed Salem, Rui Wen, Michael Backes, Shiqing Ma, and Yang Zhang. Dynamic backdoor attacks against machine learning models. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pages 703–718. IEEE, 2022.

[22] Laura Sanchez, Cinthya Grajeda, Ibrahim Baggili, and Cory Hall. A practitioner survey exploring the value of forensic tools, ai, filtering, & safer presentation for investigating child sexual abuse material (csam). *Digital Investigation*, 29:S124–S142, 2019.

[23] Carey Shenkman, Dhanaraj Thakur, and Emma Llansó. Do you see what i see? capabilities and limits of automated multimedia content analysis. *arXiv preprint arXiv:2201.11105*, 2021.

[24] Muhammad Uzair Tariq, Afsaneh Razi, Karla Badillo-Urquiola, and Pamela Wisniewski. A review of the gaps and opportunities of nudity and skin detection algorithmic research for the purpose of combating adolescent sexting behaviors. In *Human-Computer Interaction. Design Practice in Contemporary Societies: Thematic Area, HCI 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part III 21*, pages 90–108. Springer, 2019.

[25] Jossie Murcia Triviño, Sebastián Moreno Rodríguez, Daniel O Díaz López, and Félix Gómez Mármol. C3-sex: A chatbot to chase cyber perverts. In *2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, pages 50–57. IEEE, 2019.

[26] Ying Xu, Kiran Raja, Raghavendra Ramachandra, and Christoph Busch. Adversarial attacks on face recognition systems. In *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*, pages 139–161. Springer International Publishing Cham, 2022.