Comments on the Safety Tech Challenge Fund Evaluation Criteria

April 8, 2022

Thank you for the opportunity to offer feedback on the proposed evaluation criteria [18] for the UK Government's Safety Tech Challenge Fund [20]. We appreciate that REPHRAIN will offer a separate evaluation of the proposed tools for detecting and preventing child sexual abuse material (CSAM) in end-to-end encrypted (E2EE) environments.

We are researchers at Princeton University's Center for Information Technology Policy who study CSAM detection and other content moderation in E2EE environments [13, 11, 22, 23]. One of us also previously served in a U.S. government role that involved addressing human trafficking and nonconsensual pornography online. In recent research, we developed a private protocol for detecting CSAM based on perceptual hash matching [11]; our current work focuses on cryptographic transparency for private hash matching systems [22], systematizing content moderation options for E2EE environments [23], and examining the role of content moderation methods in federal CSAM prosecutions.

We write to offer specific recommendations for improving the evaluation criteria based on our expertise in and experience researching methods for CSAM detection in E2EE environments. Our comments focus on resistance to abuse, auditability, disclosure and comparison of false positive rates, transparency criteria, and considerations about security, privacy, and abuse reporting. Before turning to those substantive suggestions, we begin with overall feedback on REPHAIN's proposed criteria and process for soliciting comments.

Overall feedback on the proposed criteria and comment process. After the current comment period concludes, we urge REPHRAIN to circulate revised evaluation criteria and to offer an additional opportunity for comment before finalizing the document. The current draft of the evaluation criteria lacks specificity in the attributes of systems that REPHRAIN will evaluate and how REPHRAIN will conduct that evaluation. The existing draft criteria total about two pages of "example questions" grouped under categories of considerations, rather than a comprehensive and detailed set of rigorous evaluation criteria. Given the importance of the Safety Tech Challenge to child safety and cybersecurity policy, the need to understand how Challenge proposals will be evaluated, the expertise of REPHRAIN's participants, and the stated goals of "strict evaluation criteria" and "detailed guidance," we recommend another opportunity for public comment on a draft with more developed and specific evaluation criteria.

We also urge wider distribution of the next opportunity for comment to ensure that a broader group of expert stakeholders have fair notice and a chance to participate. The current opportunity for comment appeared predominantly through the REPHRAIN website, social media accounts, and contact list—and only gave two weeks for submission. This level of notice and opportunity to participate is below the norm for public consultation and comment processes, especially for a topic that involves significant technical complexity and public policy implications. Broader outreach and additional time are essential for meeting the evaluation's stated goals, and we recommend a further opportunity to comment that will integrate more diverse perspectives about the challenges of E2EE, content moderation, and detecting and preventing CSAM.

Resistance to abuse and conflicts of jurisdiction. One of the key concerns for E2EE content detection systems is that a foreign government may demand scanning for material that the government deems politically unacceptable but that we would consider free speech guaranteed as a human right. This perspective was repeatedly emphasized by civil society organizations and academic researchers [6, 21, 16, 7, 8, 14, 19, 10, 15, 1, 17] in response to Apple's recent proposal for a privacy-preserving CSAM detection system using sophisticated cryptographic methods [2]. A CSAM detection system is inherently a dual-use technology—

the same methods used for identifying CSAM could be applied to other types of content—and technical, legal, and operational safeguards against repurposing these systems are essential considerations.

While concerns about E2EE and content moderation have centered on demands from nondemocratic governments, similar jurisdictional issues arise for categories of content like contraband that may have subtly different rules across different jurisdictions. A system for detecting and preventing CSAM (or any other type of problematic content) in E2EE environments must account for these jurisdictional variations.

We recognize that the current Safety Tech Challenge is for proofs of concept rather than completed products. But many aspects of the draft evaluation criteria reflect forward-looking concerns, such as maintainability, compliance, accountability of human oversight, and ease of evasion. We emphatically suggest that now is also the time to start planning for the difficult and foreseeable long-term policy problems associated with systems like these. And, at minimum, if systems lack effective means of addressing these challenges REPHRAIN must document that in its evaluation.

Recommendations:

- The evaluation criteria should expressly address the policy problems posed by differing government demands about content, which may undermine free speech and human rights. These demands may not be public and may involve coercion.
- In evaluating how a system accounts for possible abusive repurposing, REPHRAIN should describe and evaluate specific technical, legal, and operational safeguards.

Auditability of CSAM detection systems. The current evaluation criteria note that, if a CSAM detection system depends on matching a database of known content, "it is crucial that such data can be audited and authenticated." We entirely agree, and we recommend broadening the point: auditability is essential for any CSAM detection system, including those based on machine learning, to ensure public confidence and reduce the risk of misuse.

Audit methods for AI models are still in their infancy, to be sure, and an audit must cover more than checking a database of content. At minimum, an auditor should be able to evaluate a model with the range of black-box methods that have been developed in academic literature and are now deployed in industry production settings.

In addition to examining auditability for a broader class of designs, we recommend that the evaluation criteria clearly delineate between different types of auditability. A system that can be audited by any user is very different from a system that can only be audited by a trusted third-party group. Similarly, a system that can be audited down to the source code implementation is very different from a system that can only be audited through black-box testing. A system that provides cryptographically verifiable proof of its properties is very different from a system that relies on the good faith and diligence of auditing personnel.

As we noted earlier, a leading concern with CSAM detection systems in E2EE environments is pressure from foreign governments to scan for content that is not CSAM. Rigorous auditability is an essential component of reducing that risk.

Recommendations:

- The evaluation criteria should expressly address auditability for all CSAM detection systems, not only systems that involve databases of known content.
- The evaluation criteria should clearly delineate between public and private audits, open-source and black-box audits, and cryptographic and process audits. The criteria should note that public, open-source, and cryptographic audits are strongly preferable.

Disclose and compare false positive rates. The draft evaluation criteria include questions about how CSAM detection false positives are "defined and measured" and "potential unintended consequences of false positives." The criteria do not, however, discuss disclosure and evaluation of the false positive rate itself. If these proof-of-concept systems move into widespread adoption in the U.K., the false positive rate will have significant implications for both scalability and user privacy. To illustrate the scale of the problem, if WhatsApp were to implement a CSAM scanner with a 1% false positive rate over all messages globally, it would result in 1 billion false positives per day [24], implying that there would likely be hundreds of thousands, or even millions, of false positives daily in the U.K. alone. If the reporting system involves human review, this poses a significant challenge for examining all false positives in a timely manner, and it also reflects a significant loss of privacy for the average user. Thus, it is of paramount importance that the

challenge entries are evaluated in part on their false positive rate. Ensuring the evaluation uses a realistic set of content is also essential, because small differences between the types of content used in evaluation and the types of content shared by E2EE users can result in significant differences in the false positive rate.

Recommendation:

- The evaluation criteria should include stating and analyzing the false positive rate, recognizing the importance of low FPR for scalability and user privacy.
- The criteria should examine how realistic the data used for system evaluation are in comparison to the data shared by E2EE users.

Clarify transparency criteria. The draft evaluation criteria call for "reasonable disclosure regarding how and when a CSAM prevention or detection system is engaging with the user, without enabling offenders to circumvent the system." The draft criteria also encourage comprehensive documentation of CSAM detection systems to "help the community to build trust." We agree with the goals of these evaluation components, and we recommend clarifying the different types of transparency that a system provides, such as transparency about design, implementation, prior evaluations, training data, matching data, what happens in the event of detection, matching results from deployment, and other relevant considerations.

We especially encourage including transparency about false positives in the evaluation criteria, since false positives could involve a compromise of privacy or other burdens for users. This transparency could be aggregate (e.g., periodic reports) or it could be individual (e.g., notifying a user).

Recommendations:

- The evaluation criteria should delineate among different types of transparency, noting the types of concerns that each could address.
- The evaluation criteria should particularly describe transparency about false positives, given the possible implications for users.

Separate concerns about security, user privacy, victim privacy, and abuse of reporting mechanisms. In the current draft evaluation criteria, concerns about security, user privacy, victim privacy, and mechanism abuse are all consolidated under "Privacy and Security." We recommend developing "Security" into a distinct evaluation category, focusing on safeguards against adversarial attempts to manipulate the operation of a system. These protections could be technical or operational in nature.

We also suggest that "Privacy" become a separate category, encompassing the related topics of user privacy and victim privacy. User privacy addresses the impact on an E2EE user who would otherwise not have their communications analyzed, disclosed, other otherwise acted upon. Victim privacy addresses the impact on a person who appears in CSAM, where the nature of a machine learning model or CSAM database may involve possible additional harm. In some instances, the user of an E2EE service may also be a CSAM victim.

The REPHRAIN draft evaluation criteria appear to include consideration of the usability, effectiveness, and abuse of user reporting mechanisms. (The document is somewhat ambiguous as to what aspects of the system should have mitigation strategies against "abuse or unintended use.") We support addressing those criteria, especially in the context of client-side detection systems that might prompt user reports. (The document is similarly ambiguous about how it will evaluate the privacy and integrity properties of "client-side scanning," which we find puzzling since a key goal of client-side scanning is often to prompt users to take action outside the E2EE environment.) We also encourage specifically addressing mistaken and abusive user reporting because of the significant risk of burden and harm associated with erroneous reports. The cryptography research literature describes several helpful protocols for ensuring the legitimacy (but not accuracy) of user reports [5, 9, 4, 25, 12, 3], including a protocol deployed at scale in Facebook Messenger [5]. These methods are the predominant focus of academic literature on E2EE and content moderation so far, but they do not appear to have been meaningfully addressed by the Challenge participants. We recommend that REPHRAIN include user reporting considerations in its evaluation and, to the extent a proposal is silent on those considerations, note as much.

Recommendations:

- The evaluation criteria should include a "Security" category.
- The evaluation criteria should include a "Privacy" category, including specific considerations for user and victim privacy.

- The evaluation criteria should address user reporting in addition to automated reporting. User reporting considerations are especially relevant to client-side detection systems, since those systems can (and often are intended to) prompt reports which have privacy and other impacts.

Respectfully submitted,

Sarah Scheffler
Postdoctoral Researcher, Center for Information Technology Policy, Princeton University

Jonathan Mayer
Assistant Professor of Computer Science and Public Affairs, Princeton University

Anunay Kulshrestha
Graduate Researcher, Center for Information Technology Policy, Princeton University

This comment reflects solely our own views.

References

[1] Hal Abelson et al. "Bugs in our Pockets: The Risks of Client-Side Scanning". In: *arXiv preprint arXiv:2110.07450* (2021). URL: https://arxiv.org/abs/2110.07450.

[2] Apple, Inc. *Expanded Protections for Children*. 2021. URL: https://www.apple.com/child-safety/.

[3] Long Chen and Qiang Tang. *People Who Live in Glass Houses Should not Throw Stones: Targeted Opening Message Franking Schemes*. Cryptology ePrint Archive, Report 2018/994. https://eprint.iacr.org/2018/994. 2018.

[4] Yevgeniy Dodis et al. "Fast Message Franking: From Invisible Salamanders to Encryptment". In: *Advances in Cryptology – CRYPTO 2018, Part I*. Ed. by Hovav Shacham and Alexandra Boldyreva. Vol. 10991. Lecture Notes in Computer Science. Santa Barbara, CA, USA: Springer, Heidelberg, Germany, Aug. 2018, pp. 155–186. DOI: 10.1007/978-3-319-96884-1_6.

[5] Facebook. *Messenger Secret Conversations Technical Whitepaper*. July 2016. URL: https://cdn.startupitalia.eu/wp-content/uploads/sites/14/2016/10/secret%5C_conversations%5C_whitepaper-1.pdf.

[6] Sharon Bradford Franklin and Greg Nojeim. *International Coalition Calls on Apple to Abandon Plan to Build Surveillance Capabilities into iPhones, iPads, and Other Products*. Aug. 2021. URL: https://cdt.org/insights/international-coalition-calls-on-apple-to-abandon-plan-to-build-surveillance-capabilities-into-iphones-ipads-and-other-products/.

[7] Daniel Kahn Gillmor. "Apple's New 'Child Safety' Plan for iPhones Isn't So Safe". In: *American Civil Liberties Union* (Aug. 2021). URL: https://www.aclu.org/news/privacy-technology/apples-new-child-safety-plan-for-iphones-isnt-so-safe/.

[8] Matthew D. Green and Alex Stamos. "Apple Wants to Protect Children. But It's Creating Serious Privacy Risks." In: *New York Times* (Aug. 2021). URL: https://www.nytimes.com/2021/08/11/opinion/apple-iphones-privacy.html.

[9] Paul Grubbs, Jiahui Lu, and Thomas Ristenpart. "Message Franking via Committing Authenticated Encryption". In: *Advances in Cryptology – CRYPTO 2017, Part III*. Ed. by Jonathan Katz and Hovav Shacham. Vol. 10403. Lecture Notes in Computer Science. Santa Barbara, CA, USA: Springer, Heidelberg, Germany, Aug. 2017, pp. 66–97. DOI: 10.1007/978-3-319-63697-9_3.

[10] Seny Kamara et al. "Outside looking in: Approaches to content moderation in end-to-end encrypted systems". In: *Center for Democracy and Technology* (2021). URL: https://cdt.org/wp-content/uploads/2021/08/CDT-Outside-Looking-In-Approaches-to-Content-Moderation-in-End-to-End-Encrypted-Systems.pdf.

[11] Anunay Kulshrestha and Jonathan Mayer. "Identifying Harmful Media in End-to-End Encrypted Communication: Efficient Private Membership Computation". In: *30th USENIX Security Symposium*. 2021, pp. 893–910.

[12] Iraklis Leontiadis and Serge Vaudenay. *Private Message Franking with After Opening Privacy*. Cryptology ePrint Archive, Report 2018/938. https://eprint.iacr.org/2018/938. 2018.

[13] Jonathan Mayer. *Content moderation for end-to-end encrypted messaging*. 2019. URL: https://www.cs.princeton.edu/~jrmayer/papers/Content_Moderation_for_End-to-End_Encrypted_Messaging.pdf.

[14] Jonathan Mayer and Anunay Kulshrestha. "We built a system like Apple's to flag child sexual abuse material — and concluded the tech was dangerous". In: *Washington Post* (Aug. 2021). URL: https://www.washingtonpost.com/opinions/2021/08/19/apple-csam-abuse-encryption-security-privacy-dangerous/.

[15] India McKinney and Erica Portnoy. "Apple's Plan to "Think Different" About Encryption Opens a Backdoor to Your Private Life". In: *Electronic Frontier Foundation* (Aug. 2021). URL: https://www.eff.org/deeplinks/2021/08/apples-plan-think-different-about-encryption-opens-backdoor-your-private-life.

[16] Nat Meysenburg et al. "A Technical Explainer on Apple's Concerning Privacy Changes". In: *New America* (Aug. 2021). URL: https://www.newamerica.org/oti/briefs/a-technical-explainer-on-apples-concerning-privacy-changes/.

[17] Nilay Patel, Riana Pfefferkorn, and Jennifer King. *Here's why Apple's new child safety features are so controversial*. Aug. 2021. URL: https://www.theverge.com/22617554/apple-csam-child-safety-features-jen-king-riana-pfefferkorn-interview-decoder.

[18] Claudia Peersman et al. *Scoping the Evaluation of CSAM Prevention and Detection Tools in the Context of End-to-end encryption Environments*. Version 1.1. Mar. 2022. URL: https://cpb-eu-w2.wpmucdn.com/blogs.bristol.ac.uk/dist/1/670/files/2022/03/E2EE_evaluation_criteria_document24.03.21.pdf.

[19] Eric Rescorla. "More on Apple's client-side CSAM scanning". In: *Educated Guesswork* (Aug. 2021). URL: https://educatedguesswork.org/posts/apple-csam-more/.

[20] Safety Tech Innovation Network. *Safety Tech Challenge Fund*. 2022. URL: https://www.safetytechnetwork.org.uk/innovation-challenges/safety-tech-challenge-fund/.

[21] Julian Sanchez. "Apple's iPhone: Now With Built-In Surveillance". In: *Cato Institute* (Aug. 2021). URL: https://www.cato.org/blog/apples-iphone-now-built-surveillance.

[22] Sarah Scheffler, Anunay Kulshrestha, and Jonathan Mayer. *Public Verification for Private Hash Matching: Challenges, Policy Responses, and Protocols*. 2022.

[23] Sarah Scheffler and Jonathan Mayer. *SoK: Content Moderation in End-to-End Encryption*. 2022.

[24] Manish Singh. "WhatsApp is now delivering roughly 100 billion messages a day". In: *TechCrunch* (Oct. 2020). URL: https://techcrunch.com/2020/10/29/whatsapp-is-now-delivering-roughly-100-billion-messages-a-day/.

[25] Nirvan Tyagi et al. "Asymmetric Message Franking: Content Moderation for Metadata-Private End-to-End Encryption". In: *Advances in Cryptology – CRYPTO 2019, Part III*. Ed. by Alexandra Boldyreva and Daniele Micciancio. Vol. 11694. Lecture Notes in Computer Science. Santa Barbara, CA, USA: Springer, Heidelberg, Germany, Aug. 2019, pp. 222–250. DOI: 10.1007/978-3-030-26954-8_8.