The Alan Turing Institute

Manchester Metropolitan University

Centre for Emerging Technology and Security

RESEARCH REPORT

Privacy-preserving Moderation of Illegal Online Content

Sam Stockwell, Georgia Wake, Tooska Dargahi, Oluwaseun Ajao, Annabel Latham, Ahmed Danladi Abdullahi and Dan Sexton

April 2025

About CETaS

The Centre for Emerging Technology and Security (CETaS) is a research centre based at The Alan Turing Institute, the UK's national institute for data science and artificial intelligence. The Centre's mission is to inform UK security policy through evidence-based, interdisciplinary research on emerging technology issues. Connect with CETaS at cetas.turing.ac.uk.

This research was supported by The Alan Turing Institute's Defence and National Security Grand Challenge. All views expressed in this report are those of the authors, and do not necessarily represent the views of The Alan Turing Institute or any other organisation.

Acknowledgements

The authors are grateful to all those who took part in a focus group for this project, without whom the research would not have been possible. The authors are also grateful to: Jamie (GCHQ); Dr Pedro Freire (Aston University); and Dr Gareth Tyson (Queen Mary University of London) for reviewing an earlier version of the report.

Cite this work as: Sam Stockwell et al., "Privacy-preserving Moderation of Illegal Online Content," *CETaS Research Reports* (April 2025).

Executive Summary

This CETaS Research Report examines promising content moderation solutions that can help social media platforms and end-to-end encrypted (E2EE) services fulfil their new legal duties to remove illegal online content under the UK Online Safety Act (OSA). It also seeks to understand what metrics can be used to better assess the effectiveness of moderation methods, as well as measure their impact on user privacy when they involve E2EE protocols.

As reflected in the real-world harm to users caused by rising volumes of illegal content disseminated across online domains, effective responses to this threat have been challenging to implement at scale. To further complicate these efforts, detecting such material on E2EE services – where only the sender and the recipient can view a message – involves a difficult balance between safeguarding users and minimising privacy intrusiveness.

Based on an extensive analysis of existing literature and focus groups with experts from different sectors, this report explores current challenges in content moderation and makes a series of recommendations for improving the privacy-preserving nature of tools, frameworks and policies involved in illegal content detection and removal processes.

Key research findings

- **There is an urgent need to combine existing content moderation techniques with more innovative methods, to combat evolving online threats.** Malicious actors have found a variety of ways to evade current detection processes on social media and E2EE services, while generative AI models create new challenges in spotting illegal material that is partially or entirely synthetic.

- **Community-driven or automated moderation processes still require an expert human in the loop to prevent illegal content slipping through to users.** Relying solely on crowd-sourced approaches could lead to issues in reaching a consensus on contested 'borderline cases' where there is strong disagreement between users, while automated systems face challenges in detecting complex harmful material or content outside their training datasets.

- **Publishing OSA risk assessments would make tech companies more accountable for the adoption of comprehensive safety measures.** Similar public schemes in the EU and Australia have revealed concerns over how tech companies

that own multiple services are not applying their most effective detection tools consistently across such platforms, as well as significant discrepancies in the metrics tech companies use to fill out transparency reports.

- **Evaluation assessments of moderation techniques must move beyond narrow technical performance metrics**. While these frameworks are important, they neglect other considerations in the implementation of moderation systems in E2EE services, including data exposure and adversarial resistance metrics.
- **Privacy and security protections are not incompatible. Cryptographic-based techniques offer promise in detecting illegal content while preserving user privacy on E2EE networks**. Privacy-enhancing technologies such as Zero-knowledge Proofs (ZKPs) and Private Set Intersection (PSI) can assist detection systems in verifying content properties without revealing the content itself.
- **While there are promising knowledge-sharing mechanisms for best practices in moderation approaches, they are highly fragmented and harm-specific.** A centralised, cross-harms knowledge hub on illegal content could help tech companies identify effective methods for countering these criminal activities.

Recommendations

Recommendations for Tech Platforms and Standards Bodies

1. **Encryption-preserving techniques**: E2EE platforms should test powerful privacy-enhancing technologies, such as ZKPs and PSI cryptographic techniques, to reduce the privacy intrusiveness of tools used to detect illegal online content. Initial trials could focus on pre-upload content screening to identify their applicability, while further evaluation will determine the feasibility of wider implementation.

2. **Layered approach to moderation**: social media and E2EE platforms should adopt a combination of effective and scalable moderation technologies trained on relevant harm types and content formats linked to their service(s). Utilising a 'tech stack' approach, hash matching should serve as a minimum baseline before layering additional moderation techniques relative to the safety risks of a given platform. Experienced human moderators would be involved in 'borderline' or highly complex cases, where automated systems struggle to determine the best outcome.

3. **Multidimensional assessment frameworks:** social media and E2EE platforms should adopt and refine the efficiency, effectiveness and privacy-intrusion metrics outlined in this report when evaluating different moderation techniques. This would help ensure that different risks to user safety, security and privacy were adequately factored into the implementation process.

4. **Comprehensive threat modelling:** E2EE platforms should consider privacy, security and safety risks when conducting threat modelling for content moderation systems, drawing inspiration from the OWASP Top Ten framework. This would support the implementation of tools for detecting illegal content and the identification of potential privacy or security risks linked to these solutions.

5. **Standardised privacy-enhancing moderation protocols:** the ISO/IEC JTC 1/SC 27 should develop new standards of protocols and interfaces for privacy-enhancing technologies specifically used in content moderation. This would improve consistency, interoperability and scalability across platforms.

Recommendations for Government and Regulators

1. **Cross-harms knowledge hub**: the UK's Department for Science, Innovation and Technology (DSIT), in partnership with Ofcom, should establish a shared cross-harms 'knowledge hub' to centralise best practice and signal sharing for content moderation between trusted industry, academic and civil-society partners. Drawing on similar proposals such as the EU Centre on Child Sexual Abuse, the institution would help tech companies prioritise cost-effective moderation approaches targeting different harms, as well as monitor trends in criminal evasion methods.

2. **Publicly available risk assessments:** the UK Government should table an amendment to Section 9 of the OSA, which deals with user-to-user services' duties to conduct risk assessments of illegal content. The amendment should require the largest tech companies falling under Category 1 and 2B to publish standardised risk assessments based on Ofcom's four-step process, thereby enhancing comparative analysis of different approaches to safety.

3. **Centralised risk assessment repository:** Ofcom should create a centralised and publicly available data repository based on OSA risk assessments submitted by social media and E2EE platforms that fall under Category 1 and 2B. Such public scrutiny would help incentivise large tech companies to adopt comprehensive safety measures and go beyond the legal minimum. Modelled on the EU Digital Service Act Transparency Database, it should incorporate standardised templates for submissions, and should be continually updated and configurable for users to select metrics of interest.

4. **Consistency in detection systems:** as part of its new OSA enforcement programme, Ofcom should ensure tech companies that own multiple services falling under Category 1 and 2B apply their most effective detection systems consistently across all platforms. This would reduce vulnerabilities in services with weaker safety protections that malicious actors could target to evade detection.

5. **Hash matching database standardisation**: the Home Office should coordinate with child safety organisations and industry partners to discuss ways to standardise the classification methods used across UK-based child sexual abuse material (CSAM) hash matching repositories (e.g. the Child Abuse Image Database). This would mitigate the challenges of comparatively analysing CSAM content shared by offenders on different platforms.

6. **Online harms landscape mapping:** Ofcom should conduct an exercise to map expert organisations across the online harms ecosystem, beyond data-rich harm types such as CSAM or terrorist and violent extremist content (TVEC). This should include those tackling material related to the 15 other priority offences listed in the OSA (e.g. suicide, human trafficking and animal cruelty) to identify a wide range of best practices in privacy-preserving content moderation.

Glossary

Category 1 and 2B platforms: legal categorisations under the UK Online Safety Act for the largest tech platforms that use content recommender systems and/or enable user-to-user services (e.g. direct messages).

Child Sexual Abuse Material (CSAM): sexually exploitative and illegal content involving children.

End-to-end encryption (E2EE): a method of encrypting data so that only the sender and the intended recipient(s) can access and decrypt the content.

Federated Learning (FL): allows machine learning models to be trained across multiple devices or servers holding local data samples without exchanging the data itself.

Hash matching: compares a piece of content against a database of illegal content through a unique identifier (hash) to determine whether there is a match.

Message Franking (MF): enables the verifiable reporting of harmful content in encrypted communications by cryptographically linking messages to their senders while preserving overall confidentiality.

Private Set Intersection (PSI): involves detecting whether user content matches a database of known harmful material without sharing either the full content or the database.

Searchable Symmetric Encryption (SSE): involves searching for keywords in encrypted text that may signify the presence of illegal material without decrypting the actual content.

Secure Multi-party Computation (SMPC): involves multiple parties jointly analysing potentially harmful content without exposing private data.

Terrorist and Violence Extremist Content (TVEC): content produced by or supportive of groups that identify as, or have been designated as, terrorist or violent organisations.

Trusted Execution Environments (TEEs): provide an isolated computational space where sensitive operations (e.g. illegal content screening) can be performed with hardware-level protection against unauthorised access.

Zero-knowledge Proofs (ZKPs): involve one party confirming the authenticity of a piece of content to another party without revealing any other information.

1. Introduction

The OSA came into force in 2023 and was designed to place greater responsibility on tech companies to protect the safety of online users in the UK.[1] Under the OSA, Ofcom can require digital platforms to remove "illegal content" that is publicly posted on their sites.[2] Illegal content is defined in the OSA as content that amounts to a criminal offence – such as TVEC and CSAM, as well as 15 other priority offences.[3]

1.1 What is content moderation?

Content moderation, as defined by Ofcom, relates to activities aimed at "removing, or reducing the visibility of, potentially harmful content."[4] As Figure 1 shows, moderation processes follow similar approaches on most platforms:

1) A pre-moderation filter (often automated) will identify whether the content being submitted matches any harmful known copies stored on a database accessible by the platform, banned keyword searches or violations of the platform's policies or the law.

2) If it is considered safe content, it is published on the site.

3) If not, the content is either deleted, blocked or altered in some way (e.g. blurred or de-ranked). The user's account may also be suspended or reported. Each of these actions depends on the confidence score produced by the automated system, alongside the severity of the material in relation to violations of the platform's policies or the law.

4) Users can appeal any of the decisions in step 3 if they disagree with the outcome.

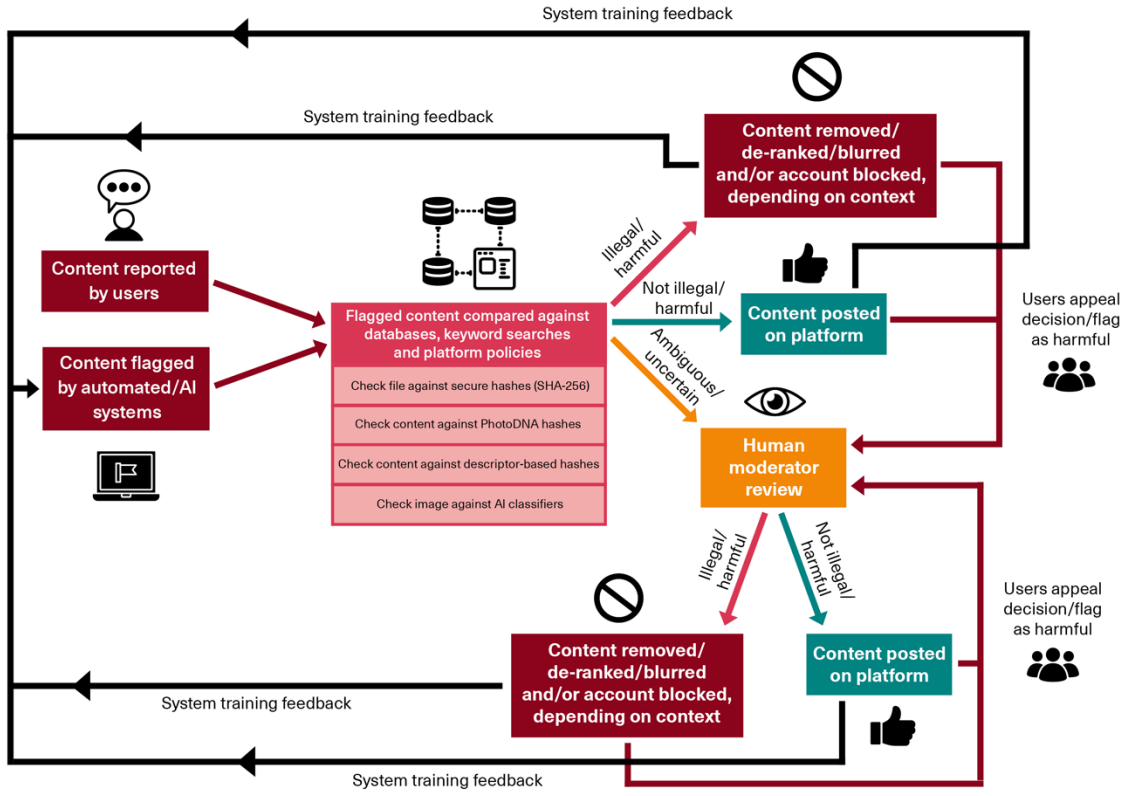
[1] "Online Safety Act 2023" (UK).
[2] Ofcom, "Time for tech firms to act: UK online safety regulation comes into force," 16 December 2024, https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/time-for-tech-firms-to-act-uk-online-safety-regulation-comes-into-force/.
[3] For the other priority offences, see: Ofcom (a), *Protecting people from illegal harms online*: *Risk Assessment Guidance and Risk Profile* (December 2024), 9, https://www.ofcom.org.uk/siteassets/resources/documents/online-safety/information-for-industry/illegal-harms/risk-assessment-guidance-and-risk-profiles.pdf?v=390984.
[4] Ofcom (a), *Content moderation in user-to-user online services* (September 2023), 3, https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/online-research/online-harms/2023/content-moderation-report.pdf?v=330128.

5) Users can also flag any content already posted (e.g. in step 2) that they believe to be harmful but that was not detected by the pre-moderation process.

6) Subsequent appeal reviews can be completed by human moderators (with or without support from automated and/or AI-based methods) to determine whether the initial decision was appropriate.

7) Data from steps 2–6 can be sent to automated moderation systems via feedback loops to refine and improve the overall process.

Figure 1: Simplified overview of content moderation processes



Source: Adapted from Cambridge Consultants, "Use of AI in Content Moderation," Ofcom, 2019, 5.

Despite widespread recognition of the need to counter the circulation of illegal online content and corresponding real-world harms, effective moderation solutions have often struggled to overcome a variety of risks and obstacles.[5] From a human rights perspective, there are tensions between balancing fundamental rights of privacy and freedom of

[5] Home Office, "Joint Statement: Tackling child sexual abuse in the age of Artificial Intelligence," 6 November 2023, https://www.gov.uk/government/publications/tackling-child-sexual-abuse-in-the-age-of-artificial-intelligence/joint-statement-tackling-child-sexual-abuse-in-the-age-of-artificial-intelligence.

expression with user safety when determining whether to remove any content.[6] Although illegal online material poses dangers to users, there is also a risk that over-moderation will have a disproportionate impact on user privacy and free speech. E2EE services pose further challenges to successful moderation. This is due to the need to both determine whether content amounts to an illegal offence and to preserve the E2EE protocols that maintain the confidentiality of law-abiding users and their messages.[7]

Since the OSA came into force, Ofcom has released a series of codes of practice and guidance documents detailing how tech companies should comply with the legislation.[8] These documents cover the causes and impacts of illegal harms; how services should assess and mitigate the risks of such harms; how services can identify illegal content; and Ofcom's approach to enforcing these measures.[9]

So far, Ofcom's guidelines have been mostly non-prescriptive regarding how individual platforms should abide by the OSA, so long as they adequately assess risks of illegal content on their own platforms and achieve the primary goal of removing such content.[10] Yet to help tech companies meet their legal obligations, it is vital to understand which solutions allow for the more effective removal of illegal material, as well as those applicable to E2EE services where there are unique risks to user privacy. Indeed, the OSA provides Ofcom with the power, where appropriate, to recommend specific technologies for platforms to detect and remove such harmful content.[11]

---

[6] Sreeprasad Govindankutty and Punit Goel, "Data Privacy and Security Challenges in Content Moderation Systems," *SSRN* (October 2024), https://dx.doi.org/10.2139/ssrn.5076831.

[7] Charles Duan and James Grimmelmann, "Content Moderation on End-to-End Encrypted Systems: A Legal Analysis," *Georgetown Law Technology Review* (January 2024), 3-9, https://georgetownlawtechreview.org/content-moderation-on-end-to-end-encrypted-systems-a-legal-analysis/GLTR-01-2024/.

[8] Ofcom, "Statement: Protecting people from illegal harms online," 24 March 2025, https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/statement-protecting-people-from-illegal-harms-online/; Ofcom (a), "Quick guide to illegal content risk assessments," 24 March 2025. https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/quick-guide-to-online-safety-risk-assessments/.

[9] Ibid.

[10] Ibid.

[11] Ofcom, *Protecting people from illegal harms online: Guidance on content communicated 'publicly' and 'privately' under the Online Safety Act* (December 2024), https://www.ofcom.org.uk/siteassets/resources/documents/online-safety/information-for-industry/illegal-harms/guidance-on-content-communicated-publicly-and-privately-under-the-online-safety-act.pdf?v=388093.

1.2 Research methodology

Within this context, this project identifies privacy-preserving techniques and policy improvements that will enable services to effectively tackle the problems of illegal content on both social media and E2EE platforms. It seeks to answer the following research questions:

- **RQ1**: What metrics can be used to assess the effectiveness and efficiency of content moderation methods?
    - **RQ1A:** How do we define, quantify and measure the impact on individual privacy of different content moderation and reporting methods?
- **RQ2:** How might tech companies improve the effectiveness and efficiency of existing content moderation policies designed to remove illegal content on social media platforms, including on encrypted data?
- **RQ3:** What existing and emerging technical capabilities should tech companies be exploring to further enhance content moderation strategies?
- **RQ4:** How can robust privacy guarantees be embedded into the aforementioned techniques to preserve user privacy with illegal content detection on E2EE networks?

Data collection for this study was conducted between September 2024 and March 2025, involving two core research activities:

1. **Literature review** covering the legal and policy aspects of content moderation responsibilities in the UK, as well as technical literature on privacy threat models and promising alternative content moderation techniques.
2. **Focus groups** designed to understand expert views on any promising content moderation techniques identified by the project team. These two sessions involved 18 experts:
    - Ten from government and regulatory bodies;
    - four from industry;
    - two from civil society; and
    - two from academia.

The project team acknowledge the following limitations of this study:

1. Our primary focus is on identifying alternative moderation policies and tools to help centralised social media platforms and E2EE services detect and remove illegal content. It is equally critical to improve other elements of the moderation workflow (e.g. policy creation, training and resourcing) and to understand the methods decentralised platforms (e.g. Mastodon) could implement to combat these harms – but both areas fall outside the scope of this report.

2. We have only considered moderation techniques that are explicitly focused on targeting *illegal* online content, as set out under the list of OSA priority offences. We have not considered other forms of content that are legal but harmful, such as misinformation – even if there may be some overlap in measures to address risks in the two categories.

3. The high-level analysis of promising moderation focuses on the broad advantages and disadvantages of such methods. We recognise that the ability to implement any of the recommended techniques will vary based on the service in question – owing to differences in resources, prevalent risk types and other platform-specific features.

1.3 Report structure

The remainder of this report is structured as follows. Section 2 describes the changing nature of the threat landscape related to illegal online content and limitations of current moderation methods. Section 3 provides an overview of different metrics for better evaluating the effectiveness, efficiency and privacy intrusiveness of content moderation techniques. Section 4 then explores ways that tech platforms can improve their moderation strategies. Finally, Section 5 details promising technical solutions for better tackling illegal online content.

2. Illegal Online Content Landscape

Although social media platforms and E2EE services have immeasurable benefits for law-abiding citizens and organisations, criminals often exploit these domains. This section describes the changing landscape of illegal online content practices, the deficiencies of current moderation solutions and the threat from generative AI models.

2.1 Online ecosystem risks

Over the past 5–10 years, developments in user accessibility and platform design in the online ecosystem have created new risks to online safety. With internet users now able to post endless amounts of content, moderation processes have become increasingly important but also increasingly strained. The scale of content posted on any given online service is now virtually impossible for human moderators to deal with alone, necessitating the use of automated solutions to prevent harmful material reaching users.[12]

Moreover, rather than only needing to cover simple text-based forum posts as in previous decades, moderation must now contend with a wide variety of content types. This includes (real and synthetic) imagery – video and livestreaming footage – and audio content, which sometimes require different approaches to detect and remove.[13] Even if one piece of viral illegal content is taken down, copies can proliferate to other platforms and cause further harm.

Corresponding to these trends, the number of victims of crime facilitated by illegal online content continues to grow exponentially. For example, the most severe types of CSAM have more than doubled since 2020, while over 300 million children under the age of 18 have been affected by such content in 2024.[14] Meanwhile, 6.6 million UK consumers lost money to online fraud content in 2024 alone.[15]

---

[12] Tarleton Gillespie, "Content moderation, AI, and the question of scale," *Big Data & Society* 7, no. 2 (August 2020), 2, https://journals.sagepub.com/doi/10.1177/2053951720943234.

[13] Robert Gorwa and Dhanaraj Thakur, *Real Time Threats: Analysis of Trust and Safety Practices for Child Sexual Exploitation and Abuse (CSEA) Prevention on Livestreaming Platforms* (Center for Democracy and Technology: November 2024), https://cdt.org/wp-content/uploads/2024/11/CDT-Research-Real-Time-Threats-hqp-final.pdf.

[14] Internet Watch Foundation, *The Annual Report 2022: #BehindTheScreens* (2022), https://www.safetolearncoalition.org/media/1286/file/IWF-Annual-Report-2022.pdf; Childlight, "Into the Light Index," https://intothelight.childlight.org/executive-summary.html.

[15] Ash Strange, "A Year On from the Online Fraud Charter," *Which?*, 18 December 2024, https://www.which.co.uk/policy-and-insight/article/a-year-on-from-the-online-fraud-charter-aefBu4h2Pre8.

An increasingly attractive channel in which to conduct such criminal activities is E2EE services, where messages can only be seen by the sender and receiver involved in a conversation.[16] These privacy features are often exploited by malicious actors because they enable offenders to disseminate illegal content anonymously, with little fear of detection by platforms or law enforcement agencies.[17] From a moderation perspective, E2EE platforms represent a particularly thorny challenge. This is due to concerns that is not technically feasible to directly access encrypted content without compromising the security and privacy of all users.[18]

While Section 5 of this paper identifies specific technical measures that can overcome these difficulties, E2EE networks should also draw inspiration from cybersecurity practices such as threat modelling when they address this problem.[19] This would help identify different risks associated with implementing moderation solutions on these networks in advance of any system deployment.[20] While most existing threat model frameworks – such as the OWASP 'Top Ten' – focus on security vulnerabilities, these could be adapted to additional safety and privacy concerns related to content moderation.[21]

For example, an E2EE platform may identify a particular online harm and want to implement corresponding safety measures to combat the risk to users. In doing so, however, it may introduce additional privacy or security threats that need to be mitigated. Likewise, measures to protect user privacy or security may introduce new safety risks that need to be addressed. By leveraging a comprehensive threat model, these issues could be mapped and then addressed through the use of the cryptographic methods listed in Section 5.

[16] Home Office, "End-to-end encryption and child safety," 20 September 2023, https://www.gov.uk/government/publications/end-to-end-encryption-and-child-safety/end-to-end-encryption-and-child-safety.
[17] Tech Against Terrorism, "Encryption: Insights from a Year of Multi-Stakeholder Discussion," January 2023, https://techagainstterrorism.org/news/2023/01/11/terrorist-use-of-end-to-end-encryption-insights-from-a-year-of-multi-stakeholder-discussion; Mar Negreiro, *E2E encryption and protection of children online* (European Parliamentary Research Service: September 2023, https://www.europarl.europa.eu/RegData/etudes/ATAG/2023/751473/EPRS_ATA(2023)751473_EN.pdf.
[18] Tech Against Terrorism (2023).
[19] CETaS focus group, 4 March 2025; Victoria Drake, "Threat Modeling," *OWASP*, https://owasp.org/www-community/Threat_Modeling/.
[20] CETaS focus group, 4 March 2025.
[21] OWASP, "OWASP Top Ten", https://owasp.org/www-project-top-ten/.

2.2 Generative AI risks

Recent advancements in AI technology have created both new benefits and new challenges for moderation processes. In particular, novel generative AI models allow users who lack technical skills to create realistic but synthetic material through simple prompts – lowering the barriers to illegal online content generation.

Generative AI is already being used in the production of sexually explicit deepfakes and CSAM content.[22] AI image generators (including so-called "nudifying apps") can create realistic CSAM, fully synthetic videos and pseudo-imagery that fall into the most severe CSAM categories.[23] This theoretically never-ending quantity of novel CSAM poses a challenge to the protection of children, since there are fears that real victims will go unnoticed as synthetic content becomes indistinguishable from real imagery.[24]

The UK Government has recognised the need to address this issue through revised legal deterrence. Although it is already illegal to possess AI-generated CSAM, new laws are aiming to target the means of illicit production. This includes new offences for possessing, creating or distributing AI tools that generate CSAM, alongside instruction manuals designed to help others do so.[25] Nevertheless, international networks of child sex offenders are still finding ways to exploit legal gaps in other jurisdictions.[26] Therefore, while it is critical to update legislation, moderation processes will also play a vital role in helping detect any illegal material intended to circumvent regulation.

Beyond CSAM, generative AI is being trialled to enhance the impact of other illegal content – such as TVEC. This includes the ability to translate extremist narratives into multiple languages for greater audience reach and converting mainstream media content into "new,

---

[22] Europol, *Internet Organised Crime Threat Assessment (IOCTA) 2024* (Publications Office of the European Union: Luxembourg), https://www.europol.europa.eu/cms/sites/default/files/documents/Internet%20Organised%20Crime%20Threat%20Assessment%20IOCTA%202024.pdf; Yiluo Wei et al., "Exploring the Use of Abusive Generative AI Models on Civitai," *MM '24: Proceedings of the 32nd ACM International Conference on* Multimedia (October 2024), https://www.eecs.qmul.ac.uk/~tysong/files/MM24-Civitai.pdf.

[23] Internet Watch Foundation, *What has changed in the AI CSAM landscape?* (July 2024), 10-18, https://www.iwf.org.uk/media/nadlcb1z/iwf-ai-csam-report_update-public-jul24v13.pdf.

[24] Ibid.

[25] Sima Kotecha, "AI-generated child sex abuse images targeted with new laws," *BBC News*, 1 February 2025, https://www.bbc.co.uk/news/articles/c8d90qe4nylo.

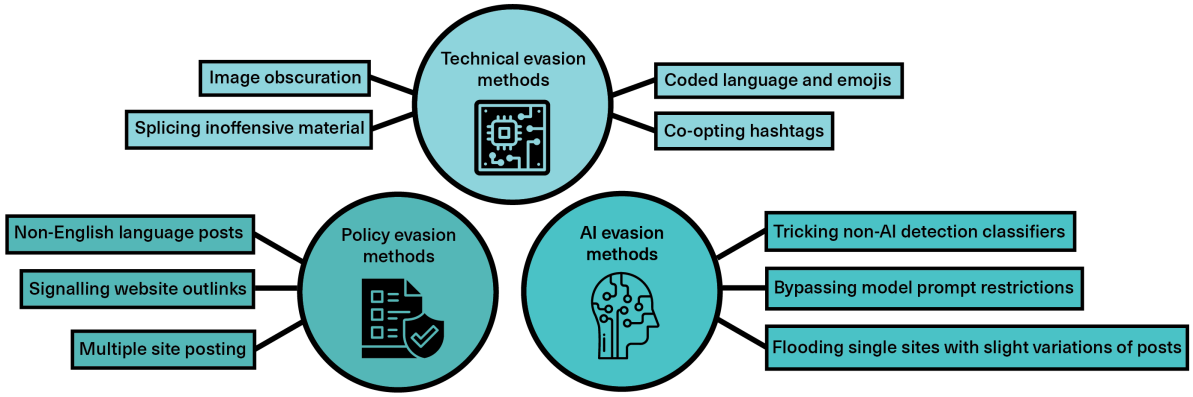
[26] Jack Burgess, "Dozens arrested in global hit against AI-generated child abuse," *BBC News*, 28 February 2025, https://www.bbc.co.uk/news/articles/czxnnzz558eo.

hate-filled versions that look and sound like the real thing."[27] There are also fears that offenders will use AI chatbots for interactive recruitment. By enhancing the personalisation of messages through large language models, terrorists could scale up messages targeted at specific demographic groups.[28]

2.3 Adversarial evasion risks

When online platforms develop or alter content moderation policies, malicious actors will seek to devise new ways to continue their criminal activities without triggering the systems in place. In other words, there is a constant race in innovation between moderation detection and evasion methods.[29] Figure 2 presents an overview of some of the most common evasion tactics offenders have used over the years.

Figure 2: Overview of common adversarial evasion tactics for content moderation



Source: Authors' analysis.

The *technical aspects* of these methods include: obscuring or altering images to evade automated image detection; using coded language or abbreviations (known as "algospeak") to evade keyword moderation; and hijacking the meaning of emojis or phrases to signify

[27] GIFCT Red Team Working Group, *Considerations of the Impacts of Generative AI on Online Terrorism and Extremism* (GIFCT: September 2023), 6, https://gifct.org/wp-content/uploads/2023/09/GIFCT-23WG-0823-GenerativeAI-1.1.pdf.

[28] Clarisa Nelu, "Exploitation of Generative AI by Terrorist Group," *International Centre for Counter-Terrorism*, 10 June 2024, https://icct.nl/publication/exploitation-generative-ai-terrorist-groups; GIFCT Red Team Working Group (2025), 7-8.

[29] Internet Watch Foundation, *How AI is being abused to create child sexual abuse imagery* (October 2023), 39, https://www.iwf.org.uk/media/q4zll2ya/iwf-ai-csam-report_public-oct23v1.pdf.

ideas to like-minded others without arousing the suspicions of platform moderators.[30] Splicing inoffensive legal material with illegal content can also bypass detection systems relying on the first few seconds of a longer form video, while the co-opting of popular hashtags helps increase the virality of shared propaganda.[31]

Malicious actors also exploit known flaws in existing *moderation policies* to their advantage. For example, Islamist terrorists often write posts in Arabic, knowing that there tends to be a lack of linguistic diversity among the moderators of even large-scale social media companies.[32] Many platforms do not monitor outlinks (e.g. hyperlinks to external websites), meaning that criminals have been known to share innocuous content but signal to like-minded users that something illegal is hosted in the link included in a post.[33] Finally, offenders may simply post the same content on multiple sites at the outset or, if moderated, later post to alternative sites or more encrypted spaces, in the hope that some of the material will evade detection.

More recently, *generative AI models* have posed further challenges to existing moderation techniques. Human traffickers, fraudsters and CSAM offenders can now generate thousands of edited versions of a single post, which can circumvent the databases of known illegal content that are used as a comparator to flag the potential sharing of new copies (hash matching).[34] Indeed, a major concern with countering AI-generated CSAM is that the influx of novel content every day requires constant updates to hash-matching databases, to ensure that these latest examples are captured.[35] Open-source AI models compound this problem, given that users can easily remove prompt restrictions designed to prevent the generation of illegal content.[36]

---

[30] Alexandra S. Levine, "From Camping To Cheese Pizza, 'Algospeak' Is Taking Over Social Media," *Forbes*, 19 September 2022, https://www.forbes.com/sites/alexandralevine/2022/09/16/algospeak-social-media-survey/; Broderick McDonald, "Extremists are Seeping Back into the Mainstream: Algorithmic Detection and Evasion Tactics on Social Media Platforms," *GNET Research*, 31 October 2022, https://gnet-research.org/2022/10/31/extremists-are-seeping-back-into-the-mainstream-algorithmic-detection-and-evasion-tactics-on-social-media-platforms/.

[31] Elisabeth Weise, "Trending hashtags co-opted by pro-terrorist accounts," *USA Today*, 11 September, 2015, https://eu.usatoday.com/story/tech/2015/09/11/pro-isis-twitter-commandeering-hijack-hashtags/72078270/.

[32] Tom Simonite, "Facebook is Everywhere; Its Moderation is Nowhere Close," *WIRED*, 25 October 2021, https://www.wired.com/story/facebooks-global-reach-exceeds-linguistic-grasp/.

[33] Tech Against Terrorism, *Terrorist Use of E2EE: State of Play, Misconceptions, and Mitigation Strategies* (September 2021), https://www.techagainstterrorism.org/hubfs/TAT-Terrorist-use-of-E2EE-and-mitigation-strategies-report-.pdf.

[34] Tech Against Terrorism, "Terrorist Use of Generative AI," https://techagainstterrorism.org/gen-ai

[35] https://www.wilsoncenter.org/article/combatting-ai-generated-csam.

[36] Internet Watch Foundation (2024), 14.

Given the rapid pace of technological development, we can expect the online content landscape to continue to evolve in the coming years – and adversaries will be quick to adopt new methods for evading detection. An agile approach is needed to ensure platforms and regulators can respond to these new threats and implement swift and robust mitigation measures.

3. Content Moderation Evaluation Metrics

Evaluating content moderation systems requires comprehensive metrics that assess both their ability to remove illegal content and their impact on user privacy. This section outlines the key metrics for measuring system performance that should be adopted when assessing moderation techniques.

3.1 Efficiency and effectiveness metrics

3.1.1 Effectiveness Metrics

Robust content moderation involves a careful balance between effectiveness (the ability to make accurate decisions) and efficiency (performance relative to the speed and scale of decisions). As platforms deal with increasing volumes of user-generated content, the need for both accurate and efficient moderation systems has become paramount. However, current evaluation assessments are often narrowly focused on technical effectiveness metrics (see Table 1). While such metrics are important, they only provide limited insights into the benefits and limitations of moderation tools.[37]

Table 1. Overview of effectiveness metrics for harmful content detection

Metric	Summary
Accuracy	The overall correctness of moderation decisions across all content types. This metric should be used in the context of content distribution and potential harms to avoid misleading outcomes.[38]
False Positive and Negative Rates	The percentage of legitimate content incorrectly flagged as harmful (false positives) and percentage of illegal content that a moderation system fails to detect (false negatives). False negatives are particularly important to prevent in this context,

[37] CETaS focus group, 5 February March 2025; CETaS focus group, 4 March 2025; Vaishali Gongane, Mousami Munot and Alwin Anuse, "Detection and moderation of detrimental content on social media platforms: current status and future directions," *Social Network Analysis and Mining* 12, no. 129 (September 2022), 35, https://link.springer.com/article/10.1007/s13278-022-00951-3.

[38] Cambridge Consultants (2019), 38.

	given the harm caused by material slipping through detection systems.[39]
Precision	The proportion of correctly identified violations among all flagged content. High precision indicates fewer false positives, reducing the risk of incorrectly removing legitimate content.[40]
Recall	The proportion of actual violations successfully identified by a moderation system. Measuring a system's ability to comprehensively find harmful content is another high-priority performance metric in the context of illegal harms.[41]
F1-Score	Combines both the ability of a moderation system to correctly identify violations (precision) and the ability to capture all violations (recall).[42]
Area Under Curve (AUC)	Measures the effectiveness of a moderation system in distinguishing between true positives (correctly identified violations) and false positives (incorrectly flagged content).[43]

3.1.2 Efficiency Metrics

Alongside effectiveness metrics, efficiency considerations can help tech companies and E2EE services understand the variables that may affect system performance on different platform types (see Table 2).

[39] CETaS focus group, 4 March 2025.
[40] Moderation API, "F1 Score", https://doi.org/10.1145/3543873.3587366.
[41] Ibid.
[42] Ibid.
[43] Pantelitsa Leonidou et al., "Privacy-Preserving Online Content Moderation with Federated Learning," in *WWW '23 Companion: Companion Proceedings of the ACM Web Conference* (April 2023), https://doi.org/10.1145/3543873.3587366.

Table 2. Overview of efficiency metrics for harmful content detection

Metric	Summary
Processing Latency	The time required for an automated detection system to analyse and moderate content.[44]
Time to Detection	The period between the posting of illegal content and an automated detection system first flagging it for review.[45]
Time to Action	The period between the posting of illegal content and the entire moderation workflow (including human reviewers) taking necessary actions in response, such as removing harmful content.[46]
Takedown Rate	The percentage of content flagged as harmful by the user or detection system that is ultimately removed. Discrepancies between flagging and takedown rates can indicate either overly sensitive detection systems or inadequate mechanisms for removal.[47]
Volume Processing Capability	An automated detection system's efficiency in processing and evaluating the volume of user-generated content within a specific time frame. It can be measured by items per second per CPU/GPU core, maximum sustainable throughput under peak load conditions and degradation patterns under stress conditions.[48]

[44] Bhatlapenumarthy and Gresham.
[45] Gideon Freud, "The Guide to Trust & Safety: Measuring Success", *ActiveFence*, 20 February 2022, https://www.activefence.com/blog/measuring-trust-and-safety/.
[46] CETaS focus group, 4 March 2025; Harsha Bhatlapenumarthy and James Gresham, "Metrics for Content Moderation", *Trust and Safety Professional Association*, https://www.tspa.org/curriculum/ts-fundamentals/content-moderation-and-operations/metrics-for-content-moderation/.
[47] Ibid.
[48] WebPurify, "Measuring the Effectiveness of Content Moderation Efforts," 7 July 2023, https://www.webpurify.com/blog/how-to-measure-content-moderation-effectiveness/.

3.1.3 User Experience Metrics

Finally, there is also a need to understand the human impact of moderation systems and their effectiveness in responding to contested removals and other user interactions (see Table 3).

Table 3. Overview of user experience metrics for harmful content moderation

Metric	Summary
Transparency Index	Measures how clearly moderation processes are explained to users. This could include transparency in process disclosure, result explanation and technical accessibility.[49]
Appeal Rate	The percentage of moderation decisions challenged by users. A high appeal rate may indicate problems with moderation quality, transparency or users' understanding of platform policies.[50]
Appeal Success Rate	The proportion of appeals resulting in a decision reversal. This metric helps identify systematic errors in moderation systems. High success rates suggest either overly aggressive initial moderation or insufficient review before takedown.[51]
User Satisfaction	Survey-based assessments of moderation fairness and effectiveness. This could include measurements of the perceived fairness of reviewer decisions, the transparency of review processes and moderators' responsiveness to user feedback.[52]
Repeat Violation Rate	The frequency with which users commit similar violations after moderation interventions, which helps evaluate deterrence

[49] Sarah Scheffler and Jonathan Mayer, "SoK: Content Moderation for End-to-End Encryption," *Proceedings on Privacy Enhancing Technologies* 2 (2023), 11-13, https://arxiv.org/abs/2303.03979.
[50] Bhatlapenumarthy and Gresham.
[51] Ibid.
[52] Centre for Data Ethics and Innovation, *The role of AI in addressing misinformation on social media platforms* (August 2021), https://assets.publishing.service.gov.uk/media/610aab37e90e0706cd12dce8/Misinformation_forum_write_up__August_2021__-_web_accessible.pdf.

	effectiveness. Effective moderation should reduce recidivism over time.[53]
Contestability Metrics	Measure how readily users can contest automated moderation decisions – including through access to appeals processes, time to appeal resolution, alternative viewpoint consideration and decision reversibility.[54]

3.2 Privacy intrusion metrics

Beyond effectiveness and efficiency, content moderation systems also involve a challenging balance between protecting the safety of users and preserving their right to privacy when implemented on E2EE services. As Section 2 highlighted, these platforms have become attractive targets for the dissemination of illegal content because law enforcement – and, sometimes, even the services themselves – are unable to access or view the encrypted content.

While Section 5 presents a series of technical solutions to address this challenge, it is also necessary to identify appropriate metrics that can ensure these methods are implemented in a way that minimises privacy intrusiveness and the risks that come with compromising E2EE protocols.[55]

This section aims to itemise and metricate the factors that should be considered when assessing the relative privacy intrusiveness of different content moderation techniques.

3.2.1 Data Exposure Metrics

The first set of privacy intrusion metrics relates to the extent to which user data is accessed, processed and retained by different actors (see Table 4).

[53] Bhatlapenumarthy and Gresham.
[54] Niva Elkin-Koren, "Contesting Algorithms: Restoring the Public Interest in Content Filtering by Artificial Intelligence," *Big Data and Society* 7, no. 2 (July 2020), https://doi.org/10.1177/2053951720932296.
[55] Scheffler and Mayer (2023), 4-5.

Table 4. Overview of data exposure metrics for harmful content moderation

Metric	Summary
Data Access Scope	The types of user content that are accessed by either the platform or law enforcement (e.g. text, images, metadata or behavioural patterns). More comprehensive access creates greater privacy risks. Access scope can be categorised across content type, access depth and access breadth.[56]
Processing Location	Where content analysis occurs (on-device, in-cloud or hybrid approaches) and which parties have exposure to the data. On-device processing is generally better than centralised analysis at preventing unauthorised parties from accessing the data, but it introduces new concerns about device integrity and user autonomy.
Retention Duration	Tracks how long flagged and unflagged content is stored for moderation purposes. Longer retention periods increase privacy risks and may conflict with data minimisation principles.
Access Frequency	Measures how often user content is analysed (continuous scanning versus triggered analysis). Continuous monitoring raises more significant privacy concerns than event-triggered analysis. Access frequency can be categorised through analysis periodicity, coverage percentage and triggering criteria.[57]
Data Minimisation Ratio	Quantifies the proportion of processed data that is necessary for moderation. This can be measured through feature extraction efficiency, processing selectivity and pseudonymisation effectiveness. [58]

[56] The Royal Society, *From Privacy to Partnership. The Role of Privacy Enhancing Technologies in Data Governance and Collaborative Analysis* (January 2023), 90-94, https://royalsociety.org/-/media/policy/projects/privacy-enhancing-technologies/from-privacy-to-partnership.pdf.
[57] Ian Levy and Crispin Robinson, "Thoughts on Child Safety on Commodity Platforms," *Cryptography and Security* (July 2022), 33-40, https://arxiv.org/pdf/2207.09506.
[58] ICO, *Privacy-Enhancing Technologies (PETs)* (June 2023), 4-5, https://ico.org.uk/media/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/privacy-enhancing-technologies-1-0.pdf.

Data Transformation Level	Assesses how data is transformed before analysis to reduce privacy risks. Transformation techniques include generalisation, suppression, perturbation and feature extraction. Higher transformation levels offer stronger privacy protections but may reduce moderation effectiveness.[59]

3.2.2 Cryptographic Protection Metrics

Alongside data exposure, moderation processes on E2EE networks require specific cryptographic safeguards. These are needed to avoid breaking the underlying architecture of the service and risking violations of user confidentiality if any flagged content is deemed legal (see Table 5).

Table 5. Overview of cryptographic protection metrics for harmful content moderation

Metric	Summary
Encryption Strength	The cryptographic method's effectiveness at protecting user data during moderation processes. This refers to the E2EE protocol's ability to maintain the full confidentiality, integrity and authentication of user messages – except in provable cases of illegal content – without introducing vulnerabilities. This evaluation must consider how cryptographic primitives are combined to provide comprehensive protection to user data across the entire moderation pipeline, not just for data at rest or in transit.[60]
Zero-knowledge Guarantees	Measure how confidently a system can analyse content without revealing unnecessary information. Strong zero-knowledge

[59] CETaS focus group, 4 March 2025; Amine Boulemtafes, Abdelouahid Derhab and Yacine Challal, "A review of privacy-preserving techniques for deep learning," *Neurocomputing* 384 (April 2020), 26, https://www.sciencedirect.com/science/article/pii/S0925231219316431.

[60] James Bartusek et al., "End-to-End Secure Messaging with Traceability Only for Illegal Content," in *Advances in Cryptology – EUROCRYPT 2023* ed. Carmit Hazay and Martijn Stam (Cham: Springer, 2023), https://link.springer.com/chapter/10.1007/978-3-031-30589-4_2.

	properties ensure that moderation systems cannot learn additional information beyond what is strictly required.[61]
Protocol Privacy Leakage	The degree to which information is inadvertently revealed through protocol design or implementation. Even cryptographically secure systems may leak information through their structure or operation.[62]

3.2.3 Adversarial Resistance Metrics

Any moderation system implemented on E2EE services to scan for potential illegal content could also introduce 'backdoor' vulnerabilities, in which malicious actors seek to exploit the system's methods to steal personal data or conduct other criminal activity.[63] Consequently, it is crucial to determine such solutions' resistance to privacy attacks and unauthorised data access (see Table 6).

Table 6. Overview of adversarial resistance metrics for harmful content moderation

Metric	Summary
Inference Attack Resistance	The moderation systems' ability to protect against unauthorised attempts to access user information. This can be measured through reconstruction accuracy under optimal attacks, information leakage quantification and membership inference vulnerability.[64]
Side-channel Leakage	The degree to which information is inadvertently exposed through timing, pattern or operational characteristics. This can be

[61] Sarah Scheffler, Anunay Kulshrestha, and Jonathan Mayer, "Public Verification for Private Hash Matching," *2023 IEEE Symposium on Security and Privacy* (May 2023), 258, https://doi.org/10.1109/SP46215.2023.10179349.
[62] Scheffler and Mayer (2023), 17.
[63] Seny Kamara et al., *Outside Looking In: Approaches to Content Moderation in End-to-End Encrypted Systems* (Center for Democracy and Technology: 2021), 15, https://cdt.org/wp-content/uploads/2021/08/CDT-Outside-Looking-In-Approaches-to-Content-Moderation-in-End-to-End-Encrypted-Systems-updated-20220113.pdf.
[64] Yiqing Hua et al., "Increasing Adversarial Uncertainty to Scale Private Similarity Testing," in *Proceedings of the 31st USENIX Security Symposium, Security 2022* (August 2022), https://www.usenix.org/system/files/sec22-hua.pdf.

	evaluated according to timing attack vulnerability, power analysis resistance and network traffic pattern obfuscation.[65]
Differential Privacy Guarantees	Measure whether specific details about individuals and their identity (ε-value) are protected during the analysis of potentially illegal content when differential privacy systems are used. Lower ε-values indicate stronger privacy protection but may suggest lower moderation utility.
Database Reconstruction Immunity	Resistance to attackers' attempts to reconstruct original content from content hashes or fingerprints.[66]
Database Integrity	The security and reliability of illegal content databases relative to insider and outsider threats. This includes the capability to ensure the accuracy, consistency and protection of data from unauthorised modifications. Threat protections can be assessed through internal quality assurances, system logs, independent audits and robust cybersecurity practices (e.g. access controls, encryption and regular vulnerability assessments).

3.2.5 Regulatory Compliance Metrics

Finally, the privacy-preserving implementation of moderation systems can be enhanced through alignment with relevant legal principles and frameworks (see Table 7).

Table 7. Overview of regulatory compliance metrics for harmful content moderation

Metric	Summary
GDPR Alignment Score	Measures adherence to UK General Data Protection Regulation (GDPR) principles. Compliance can be evaluated across key

[65] Ileana Buhan et al., "SoK: Design Tools for Side-Channel-Aware Implementations," *Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security* (May 2022), https://dl.acm.org/doi/10.1145/3488932.3517415.

[66] Sophie Hawkes et al., "Perceptual Hash Inversion Attacks on Image-Based Sexual Abuse Removal Tools," *IEEE Security and Privacy* (November 2024), 7-8, https://doi.org/10.1109/MSEC.2024.3485497.

	principles such as lawfulness, fairness, transparency, purpose limitation and data minimisation.[67]
Data Protection Impact Assessment (DPIA) Coverage	Evaluates the comprehensiveness of privacy risk assessments. DPIAs should consider the necessity and proportionality of processing; the impact on data subject rights; and security measures. For illegal content moderation, this should address the special category data implications and the additional safeguards that are implemented.[68]
Legitimate Interest Assessment	Evaluates the balancing of platform moderation interests against user privacy rights. This should include assessment of purpose specification clarity, necessity demonstration and balancing test comprehensiveness.[69]

To our knowledge, there is no standardised approach to measuring the relative privacy intrusiveness of content moderation practices across different platforms. By adopting the metrics proposed in this section, platforms could provide clear assurance to both regulators and users that all reasonable steps were taken to prevent the spread of illegal content while maintaining user privacy.

[67] ICO, "Principles and definitions", https://ico.org.uk/for-organisations/advice-for-small-organisations/frequently-asked-questions/principles-and-definitions/.
[68] ICO, "Data Protection Impact Assessments (DPIAs)", https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/accountability-and-governance/data-protection-impact-assessments-dpias/.
[69] ICO, "Legitimate interest assessment (LIA)", https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/accountability-and-governance/accountability-framework/records-of-processing-and-lawful-basis/legitimate-interest-assessment-lia/.

4. Improving Moderation Strategies

The continuous improvement of content moderation strategies is essential, given that malicious actors are constantly devising new ways to disseminate illegal content and evade existing approaches. This section provides an overview of the strategies that tech companies should adopt to reduce the risk of illegal material proliferating on their sites.

4.1 Platform policies and systems

Social media companies and E2EE platforms have frequently adapted their content moderation policies based on perceived public and political appetite.[70] In recent years, such policies have shifted towards greater protections of freedom of speech and user privacy.[71] This has included moves to increase the rollout of E2EE protocols on messaging apps, as well as transitions away from moderation solutions that incorporate human moderators and towards completely community-driven or automated approaches.[72] For example, X removed more than 80% of its trust and safety team in 2022. Meta, Google, Amazon and Discord followed suit by downsizing the number of human moderators within their organisations.[73]

Integrating automated or decentralised methods into moderation processes has certain benefits. Entirely automated pipelines can increase the scalability and speed of platforms' efforts to detect harmful content, while avoiding the need for humans to review potentially distressing material. Similarly, community-based approaches can help reduce individual

[70] Tamar Mitts, "Content moderation is a policy problem, not just a platform problem", *Princeton University Press*, 11 March 2025, https://press.princeton.edu/ideas/content-moderation-is-a-policy-problem-not-just-a-platform-problem.
[71] Liv McMahon, Zoe Kleinman & Courtney Subramanian, "Facebook and Instagram get rid of fact checkers," *BBC News*, 7 January 2025, https://www.bbc.co.uk/news/articles/cly74mpy8klo; Centre for Data Ethics and Innovation (2021); Gillespie (2020).
[72] Chris Vallance, "Facebook and Messenger to automatically encrypt messages," *BBC News*, 7 December 2023, https://www.bbc.co.uk/news/technology-67646047; McMahon et al., (2025); Nurudeen Akewushola, "Musk explains how X corrects inaccurate posts with community notes," *FactCheckHub*, 4 November 2023, https://factcheckhub.com/musk-explains-how-x-corrects-inaccurate-posts-with-community-notes/.
[73] Vittoria Elliot, "Elon Musk's Twitter Takeover Set Off a Race to the Bottom," *WIRED*, 5 November 2024, https://www.wired.com/story/elon-musk-trust-safety-industry/; Daria Dergacheva, "Platforms overwhelmingly use automated content moderation, first DSA transparency reports show," *Lab Platform Governance, Media and Technology*, 8 November 2023, https://platform-governance.org/2023/platforms-overwhelmingly-use-automated-content-moderation-first-dsa-transparency-reports-show/.

biases and 'majority rule' in decision-making.[74] Nevertheless, the complete removal of an expert human in the loop also comes with risks for countering illegal content.[75]

Community-driven moderation can suffer from the difficulty of reaching a consensus among users as to whether certain types of content should be flagged for removal, and tends to focus on harmful but legal content, such as political misinformation.[76] Furthermore, rather than eradicating bias, automated solutions risk replicating human prejudices – with a disproportionate effect on marginalised groups.[77] Automated tools can also lack contextual nuance and can struggle to detect new types of illegal content that were absent from their training data.[78] Indeed, Meta has acknowledged that its new policy of replacing human fact-checkers with AI-based techniques could allow more harmful content to appear on the platform.[79]

Owing to the deficiencies of these different approaches in isolation, social media platforms should adopt a hybrid model that leverages the benefits of both human experts and automation. This would involve deploying effective and scalable combinations of moderation technologies in a 'tech stack' – comprising layers of additional systems that,

---

[74] Peter Suciu, "Just The Facts – Are Community Notes Working On Social Media?," *Forbes*, https://www.forbes.com/sites/petersuciu/2025/03/24/just-the-facts--are-community-notes-working-on-social-media/.

[75] CETaS focus group, 5 February 2025; Gillespie (2020); Jess Brough, "Content moderation offers little actual safety on Big Social Media," *New Scientist*, 12 March 2025, https://www.newscientist.com/article/mg26535342-200-content-moderation-offers-little-actual-safety-on-big-social-media/; Yannis Theocharis et al., *Content Warning: Public Attitudes on Content Moderation and Freedom of Expression* (Content Moderation Lab: 2025), 9, https://tumthinktank.de/wp-content/uploads/ContentWarning_Report_2025_CML.pdf.

[76] Centre for Countering Digital Hate, *Rated not Helpful: How X's Community Notes system falls short on misleading election claims* (October 2024), https://counterhate.com/wp-content/uploads/2024/10/CCDH.CommunityNotes.FINAL-30.10.pdf; Will Oremus, Trisha Thadani and Jeremy B. Merrill, "Elon Musk says X users fight falsehoods. The falsehoods are winning," *The Washington Post*, 30 October 2024, https://www.washingtonpost.com/technology/2024/10/30/elon-musk-x-fact-check-community-notes-misinformation/.

[77] Andrea Stockinger, Svenja Schäfer and Sophie Lecheler, "Navigating the gray areas of content moderation: Professional moderators' perspectives on uncivil user comments and the role of (AI-based) technological tools," *New Media & Society* 27, no. 3 (August 2023), 1230, https://journals.sagepub.com/doi/full/10.1177/14614448231190901; Michael Barnes, "Online extremism, AI, and (Human) Content Moderation," *Feminist Philosophy Quarterly* 8, no. 3/4 (2022), 21, https://ojs.lib.uwo.ca/index.php/fpq/article/view/14295.

[78] Stockinger et al., (2023), 1228.

[79] Clare Duffy, "Meta is getting rid of fact checkers. Zuckerberg acknowledged more harmful content will appear on the platforms now," *CNN*, January 7 2025, https://edition.cnn.com/2025/01/07/tech/meta-censorship-moderation/index.html.

based on OSA risk assessments, would be suited to the safety risks of a given platform.[80] In other words, higher-risk services would require a more robust and multilayered tech stack. Experienced human moderators would then be involved in 'borderline' or highly complex cases, helping improve user trust in decisions where automated systems struggle to determine the best course of action or could make errors that pose serious risks to human rights (e.g. benign content incorrectly flagged and reported as illegal).[81]

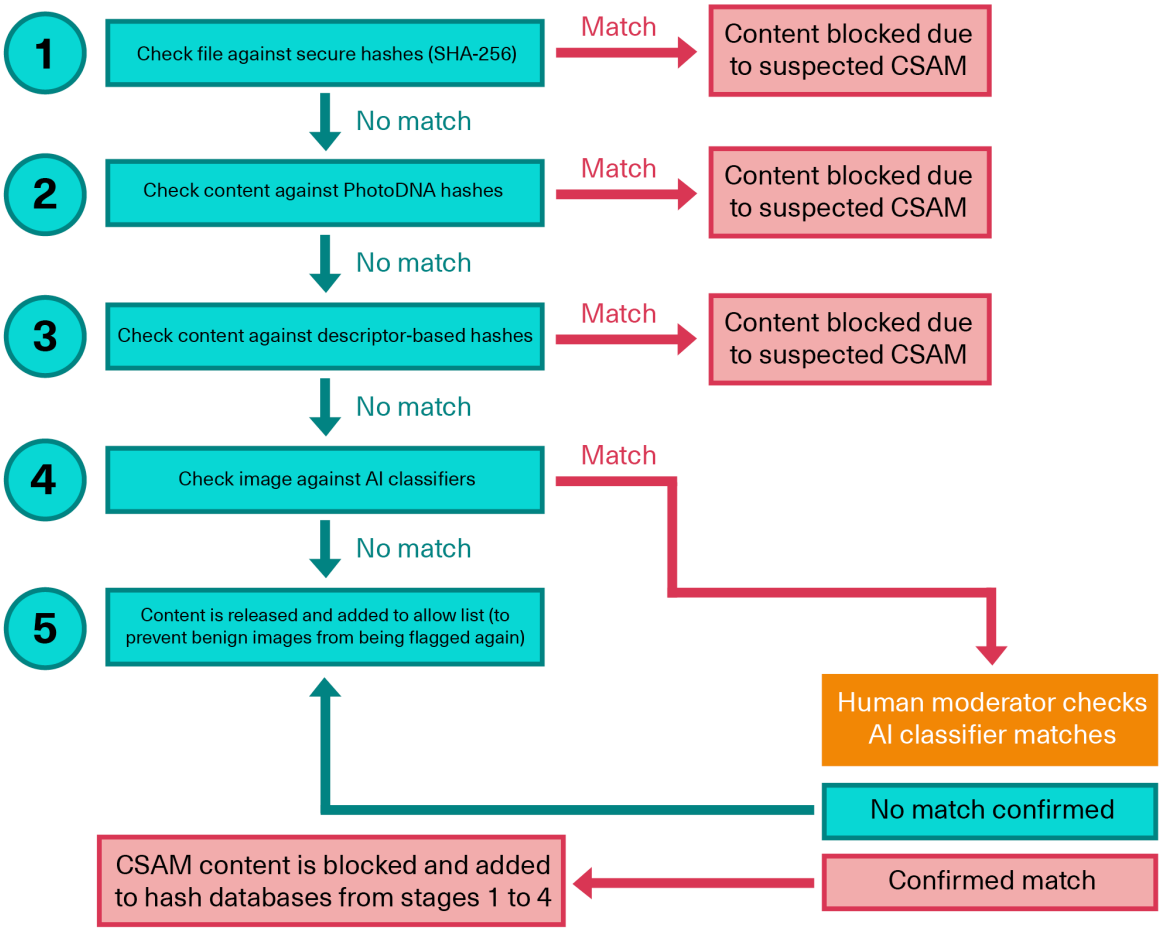
Through this process, services would start with the most accurate and efficient automated techniques before moving through the stack, as content was either blocked or released based on allow/deny listing at each stage. Human reviewers would check any flagged matches against AI classifiers, due to the risk of false positives. Figure 3 provides an example of how such a layered approach would work in practice with screening for CSAM, in which encryption-based techniques (see Section 5.2) could be combined with stages 1–4 of the process to preserve user privacy without reducing detection effectiveness. Critically, users should be given the option to appeal different blocks at any time, while feedback mechanisms from moderation decisions should be used to improve the continuous learning and refinement of automated systems.[82]

[80] CETaS focus group, 5 February 2025; Ofcom (2024a), 18-26; Joan Donovan, "Navigating the Tech Stack: When, Where and How Should We Moderate Content?," *Center for International Governance Innovation*, 28 October 2019, https://www.cigionline.org/articles/navigating-tech-stack-when-where-and-how-should-we-moderate-content/.

[81] Thiago Dias Oliva, "Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression," *Human Rights Law Review* 20, no. 4 (December 2020), 639-640, https://academic.oup.com/hrlr/article/20/4/607/6023108; Christina Pan et al., "Comparing the Perceived Legitimacy of Content Moderation Processes: Contractors, Algorithms, Expert Panels, and Digital Juries," *Proceedings of the ACM on Human-Computer Interaction* 6 (April 2022), 22, https://dl.acm.org/doi/10.1145/3512929.

[82] CETaS focus group, 4 March 2025; Maria D Molina and S Shyam Sundar, "When AI moderates online content: effects of human collaboration and interactive transparency on user trust," *Journal of Computer-Mediated Communication* 27, no. 4 (July 2022), 1, https://academic.oup.com/jcmc/article/27/4/zmac010/6648459.

Figure 3: Overview of a layered moderation process for CSAM matches



Source: Authors' analysis.

It is vital that, on top of adopting a layered moderation process, tech companies that own multiple services deploy their most effective detection tools consistently across these services. This holistic approach to implementation is needed because malicious actors target sites with weaker or fewer mechanisms in place to filter illegal content. However, a recent investigation by Australia's eSafety Commissioner found that some of the largest tech companies could be doing more in this respect. For example, while some are limiting hash matching technology to known TVEC, they have not extended this to new material of the same kind.[83] Likewise, organisations are not adopting newer and more effective hash matching tools that are available on certain services.[84]

[83] Australia eSafety Commissioner, "eSafety report reveals serious gaps in how tech industry is tackling terror and violent extremism," 6 March 2025, https://www.esafety.gov.au/newsroom/media-releases/esafety-report-reveals-serious-gaps-in-how-tech-industry-is-tackling-terror-and-violent-extremism.
[84] Ibid.

Consequently, Ofcom should ensure that tech companies that own multiple services falling under Category 1 and 2B of the OSA are consistently applying their most effective moderation tools across their platforms, as part of its new 'enforcement programme.'[85] This would reduce vulnerabilities in services that malicious actors could target to evade detection, while enhancing user safety in accordance with legal obligations.

4.2 Knowledge-sharing mechanisms

Despite many similar evasion techniques described in Section 2 happening across multiple platforms, as well as across multiple harm types, information about these methods and ways to combat them is rarely shared between platforms – except when platforms actively engage with civil society organisations to combat specific criminal activity.[86]

Some of these organisations, such as the Global Internet Forum to Counter Terrorism (GIFCT), offer mentorship services that enable members to seek advice on how to strengthen their moderation policies in line with best practices.[87] However, engagement with these bodies often involves strict membership criteria that many platforms do not meet. Other initiatives – such as the Tech Coalition's 'Lantern' programme and Robust Open Online Safety Tools – are also valuable in allowing platforms to share signals about illegal activity and open-source tools that enhance moderation processes.[88] Nevertheless, they are mostly limited to specific harms (e.g. CSAM) and primarily involve representatives from industry.

Given these challenges, DSIT and Ofcom should establish a new cross-harms 'knowledge hub' to centralise best practice and signal sharing for content moderation between trusted industry, academic, and civil society partners. This would help tech platforms prioritise cost-effective moderation approaches targeting multiple harm types, and monitor trends in malicious actors' detection evasion methods, which are currently fragmented.[89] The UK could draw inspiration from the EU's similar proposal for a Centre on Child Sexual Abuse,

[85] Ofcom, "Enforcing the Online Safety Act: Scrutinising illegal harms risk assessments," 3 March 2025, https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/enforcing-the-online-safety-act-scrutinising-illegal-harms-risk-assessments/.
[86] CETaS focus group, 5 February 2025.
[87] GIFCT, "Membership," https://gifct.org/membership/.
[88] Sean Litton, "Announcing Lantern: The First Child Safety Cross-Platform Signal Sharing Program," *The Tech Coalition*, 7 November 2023, https://www.technologycoalition.org/newsroom/announcing-lantern; Cristina Martinez, "ROOST: A Collaborative Effort for AI-Driven Online Safety," *Medium*, 12 March 2025, https://medium.com/nexstudent-network/roost-a-collaborative-effort-for-ai-driven-online-safety-42001150d45b.
[89] CETaS focus group, 5 February 2025.

which is envisaged to act as a hub of expertise and provide reliable information on identified CSAM material for swifter law enforcement responses.[90]

4.3 Illegal content databases

Hash matching databases are offered by organisations such as the Internet Watch Foundation and the GIFCT to help member platforms remove CSAM and TVEC respectively. These repositories allow tech companies to train their detection systems on a wide range of historical illicit material, with the aim of strengthening their effectiveness at flagging similar copies posted on their sites. Ofcom guidance emphasises the importance of these databases in helping combat illegal content, recommending that all services implement CSAM hash matching techniques on their platforms.[91]

However, it is also vital that existing repositories containing material from the same harm types incorporate standardised metrics. Currently, CSAM databases in the UK (e.g. the Child Abuse Image Database) have distinct labelling practices.[92] This makes it difficult to combine them for comparative analysis that could help law enforcement identify strategic trends in criminal behaviour.[93]

Ofcom should, therefore, convene child safety organisations and relevant government departments (e.g. the Home Office) to develop standardised classification methods for all UK-owned CSAM databases. This could draw on international initiatives such as INHOPE's Global Standard project, which seeks to harmonise the terminology used to classify CSAM and create an interoperable global CSAM hash set.[94]

Outside data-rich areas such as CSAM and TVEC, there is also a need to gather data on a wider range of illegal content types proscribed under the OSA. This could help reduce the risk that users will be exposed to harmful material on animal cruelty, firearms, suicide,

[90] DG HOME, "Legal framework to protect children," https://home-affairs.ec.europa.eu/policies/internal-security/protecting-children-sexual-abuse/legal-framework-protect-children_en.
[91] Ofcom, *Protecting people from illegal harms online Volume 2: Service design and user choice* (16 December 2024), https://www.ofcom.org.uk/siteassets/resources/documents/online-safety/information-for-industry/illegal-harms/volume-2-service-design-and-user-choice.pdf?v=388720; GIFCT, "GIFCT's Hash-Sharing Database," https://gifct.org/hsdb/.
[92] CETaS focus group, 5 February 2025; Home Office, "Child abuse image database," 15 May 2024, https://www.gov.uk/government/publications/child-abuse-image-database.
[93] CETaS focus group, 5 February 2025.
[94] Safe Online, "A universal language for CSAM classification," https://safeonline.global/a-universal-language-for-csam-classification-inhope/.

human trafficking and other themes.[95] As such, Ofcom should conduct an exercise to map prominent experts and organisations across the online harms ecosystem, identifying wider best practices in privacy-preserving content moderation.

4.4 Transparency reporting

Transparency reporting provides external stakeholders with data on different tech platforms' moderation decisions – and, accordingly, potential insights into how the platforms' processes work, the types of harmful content they remove and the decisions they make in response to user appeals. Despite these benefits, existing transparency reporting also suffers from several issues.

When transparency reporting is made optional for platforms, there is sometimes an incentive to *avoid* as much detail as possible. While this can be due to concerns about helping malicious actors circumvent moderation strategies, it can also protect the reputation of businesses by reducing the risk of greater criticism of their moderation decisions.[96] Additionally, databases that store historical transparency reports often reveal significant discrepancies in the metrics platforms use. These discrepancies make it harder to conduct comparative analysis that helps external stakeholders scrutinise tech companies' moderation processes.[97]

With the introduction of the OSA, tech platforms falling under the scope of the legislation are now legally required to conduct "risk assessments." Such exercises are designed to identify the risks associated with illegal content on their services and the safety measures they need to put in place to protect users.[98] Although risk assessments could help hold platforms accountable for their decisions, there is no requirement to publish the reports.[99] This risks making it more difficult for researchers, academics and others beyond Ofcom to increase scrutiny of, and compare approaches between, services.

---

[95] Ofcom (2024a), 9.

[96] CETaS focus group, 5 February 2025; Evelyn Douek, "Governing Online Speech: From 'Posts-As-Trumps' to Proportionality and Probability," *Columbia Law Review* 121, no. 3 (August 2020), 828, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3679607.

[97] Amaury Trujillo, Tiziano Fagni and Stefano Cresci, "The DSA Transparency Database: Auditing Self-reported Moderation Actions by Social Media," *Proceedings of The 28th 2025 ACM Conference on Computer-Supported Cooperative Work and Social Computing* (February 2025), 17-20, https://arxiv.org/pdf/2312.10269v4.

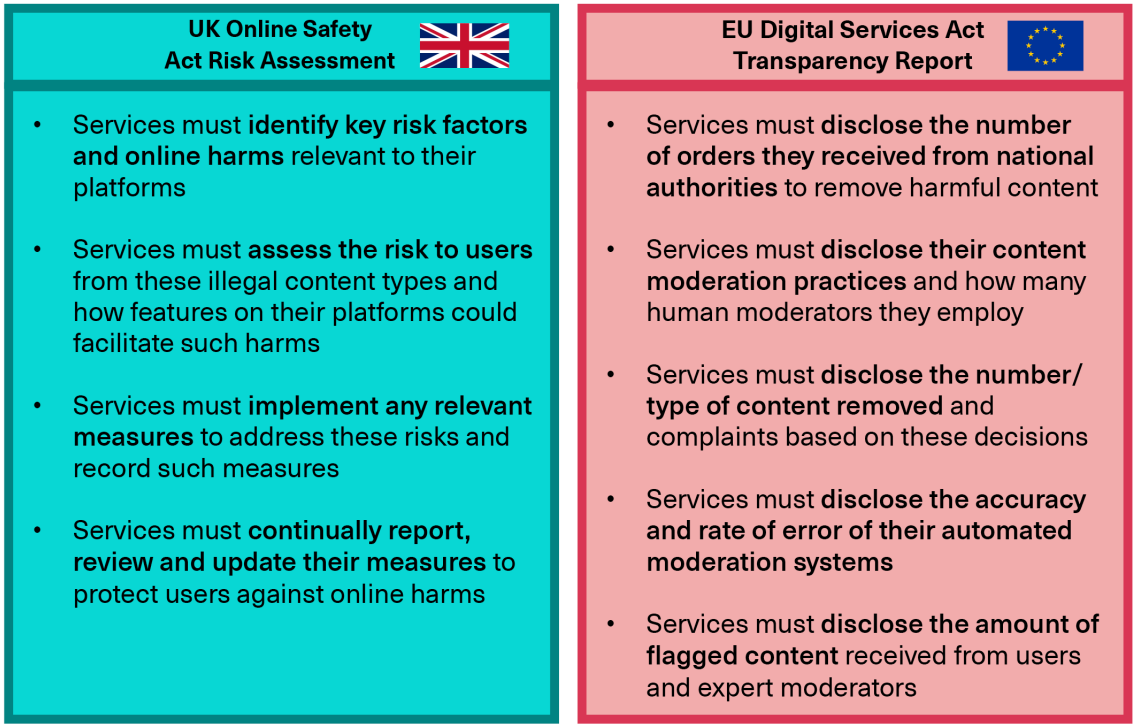
[98] Ofcom (2025a).

[99] CETaS focus group, 5 February 2025.

Accordingly, the UK Government should improve transparency and accountability by tabling an amendment to Section 9 of the OSA that addresses user-to-user services' duties to conduct illegal content risk assessments.[100] The amendment should require tech companies falling under Category 1 and 2B to publish standardised risk assessments in a way that supports comparative analysis of approaches to safety. Such risk assessments should be based on Ofcom's four-step process and should include details as to why any online harm mitigation is necessary and proportionate.[101]

Although Ofcom's risk assessment process does not provide information on moderation decisions, the EU requires these details through the Digital Services Act (DSA) (see Figure 4 for comparison). This includes legal obligations for periodic reports around user appeals, the content that was removed and reason(s) for doing so, and the use of automation in the moderation process.[102] Additionally, the DSA Transparency Database provides an open and centralised repository of these decisions, to enhance platform transparency.[103]

Figure 4: Comparison of UK OSA and EU DSA transparency reporting requirements



Source: Ofcom (2024a), 15-17; EU Commission (2024).

[100] "UK Online Safety Act" (2023).
[101] Ofcom (2025a), 15.
[102] Trujillo et al., (2025), 2.
[103] EU Commission, "DSA Transparency Database," https://transparency.dsa.ec.europa.eu/.

The EU's approach to transparency reporting provides an unprecedented volume of data to track, scrutinise and compare real-world moderation actions across different platforms – particularly with the recent introduction of standardised formats and reporting periods.[104] At the same time, introducing new transparency reporting requirements in the UK could create a "disconnected web" of platforms producing different assessments and add unnecessary workload for services or those seeking to hold them accountable.[105]

To leverage the benefits of pre-existing transparency reporting regimes while making them more effective, Ofcom should create a centralised and publicly available data repository based on risk assessments submitted by social media and E2EE platforms that fall under Category 1 and 2B of the OSA.[106] Based on a model similar to the EU's DSA Transparency Database, it should incorporate standardised templates for submissions and be configurable to allow users to select relevant metrics of interest.

[104] EU Commission, "Commission harmonises transparency reporting rules under the Digital Services Act," 4 November 2024, https://digital-strategy.ec.europa.eu/en/news/commission-harmonises-transparency-reporting-rules-under-digital-services-act; Trujillo et al., (2025).
[105] VOX-Pol, "Content Moderation, Transparency (Reporting) and Human Rights," 28 July 2021, https://voxpol.eu/content-moderation-transparency-reporting-and-human-rights/.
[106] CETaS focus group, 5 February 2025.

5. Privacy-preserving Moderation Solutions

One of the most challenging aspects of identifying moderation solutions to implement on platforms, particularly E2EE environments, is in understanding which designs provide strong effectiveness and efficiency guarantees while minimising risks to user privacy. This section sets out a series of promising moderation solutions, outlining their benefits and limitations.

5.1 Prioritisation table

When evaluating promising privacy-preserving content moderation and detection techniques that can be implemented in E2EE networks, a variety of criteria need to be considered:

1. **Effectiveness** – How accurately the detection technique identifies illegal content while minimising both false positives (incorrectly flagging legitimate content) and false negatives (missing harmful content).

2. **Efficiency** – The method's computational resources requirements, processing time and scalability across platforms of different sizes.

3. **Privacy protection** – The solution's level of intrusion into user data and communications in relation to different threat models.

4. **Technical feasibility** – The technique's current implementation readiness, deployment challenges and compatibility with existing systems.

Based on these factors, social media and E2EE platforms should test the techniques in Table 8 in controlled environments before considering wider implementation.[107]

[107] It should be noted that all these solutions will rely on correct policy formulation and dataset training practices prior to testing, though this is out of the scope of this report. For further information on the privacy-enhancing technologies discussed in this section, see: George Balston, Marion Oswald, Alexander Harris and Ardi Janjeva, "Privacy and Intelligence: Implications of Emerging Privacy Enhancing Technologies for UK Surveillance Policy," *CETaS Research Reports* (July 2022), https://cetas.turing.ac.uk/publications/privacy-and-intelligence.

Table 8. Overview of promising privacy-preserving content moderation and detection solutions

Solution	Summary	Benefits	Limitations
AI Image-to-text Moderation	Converts images into text formats for moderation purposes.	Reduces moderators' exposure to potentially harmful content. Could enable sharing databases of text linked to confirmed illicit imagery for detection system training.	Reduces intrusion into user privacy by not requiring access to the image itself, but there is uncertainty over whether these methods work in E2EE environments. Limited to image formats.
Anonymous Blocklisting	Allows recipients to block unwanted senders without revealing their identities.	Can help filter messages from blocked senders while preserving anonymity. Platforms can enforce blocking without identifying the users involved.	Introduces computational overheads. Reactive rather than proactive approach that may lead to initial harm from content exposure. Removal of user metadata makes it harder to identify malicious actors.
Federated Learning	Allows machine learning (ML) models for content moderation to be trained across multiple devices or servers holding local data samples	Enables personalised moderation while preserving privacy. Minimises the risk of centralised breaches. Implemented through research.[108]	Vulnerable to data poisoning attacks. Introduces communication overheads. Requires sufficient data on harm types to work

[108] Leonidou et al., (2023).

	without exchanging the data itself.		effectively, due to its reporting approach.
Homomorphic Encryption	Involves analysis of potentially harmful encrypted content without requiring decryption.	Preserves privacy while being highly secure. Enables E2EE moderation without creating attack vectors for misuse. Implemented through research.[109]	Computationally expensive. Limited to basic operations. Unsuitable for real-time analysis. May reduce transparency in moderation decisions by making it harder for service providers or users to understand why content was flagged.
Message Franking	Enables the verifiable reporting of harmful content in encrypted communications by cryptographically linking messages to their senders while preserving overall confidentiality.	Ensures verifiable reporting while maintaining E2EE. Preserves the privacy of honest users who report harmful material. Implemented on Facebook Messenger.[110]	Only works on user-reported content, so does not prevent further distribution if resent as a new message. Requires recipient participation.

[109] Tengfei Zheng et al., "Inspecting End-to-End Encrypted Communication Differentially for the Efficient Identification of Harmful Media," *IEEE Transactions on Information Forensics and Security* 18 (2023), https://doi.org/10.1109/TIFS.2023.3315067.

[110] Paul Grubbs, Jiahui Lu and Thomas Ristenpart, "Message Franking via Committing Authenticated Encryption," in *Advances in Cryptology – CRYPTO 2017*, ed. by Jonathan Katz and Hovav Shacham (Cham: Springer, 2017), 66-97, https://link.springer.com/chapter/10.1007/978-3-319-63697-9_3.

Private Set Intersection	Involves detecting whether user content matches a database of known harmful material without sharing either the full content or the database.	Secure matching without the exposure of non-matching data. Addresses weaknesses of perceptual hashing. Compatible with E2EE protocols.	Some designs can be computationally expensive (e.g. Oblivious Polynomial Evaluation). Limited to known content. Involves complex protocols.
Searchable Symmetric Encryption	Involves searching for keywords in encrypted data without decrypting it.	More efficient than HE. Supports verifiable reporting. Implemented through research.[111]	Limited to keyword and pattern matching. Search patterns may be leaked. Limits contextual understanding of content.
Secure Multi-party Computation	Involves multiple parties jointly analysing potentially harmful content without exposing private data.	Provides strong privacy guarantees. Allows for flexible computations. Implemented through research.[112]	High communication overheads. Complex implementation. Creates potential side-channel risks.
Trusted Execution Environments	Involve an isolated computational space where sensitive operations can be performed with hardware-level protection against	Enable platforms to run sophisticated detection algorithms in the cloud while maintaining strong privacy guarantees.	Require specialised hardware support. Create potential side-channel risks.

[111] Scheffler and Mayer (2023).

[112] Sergey Zapechnikov, "Secure Multi-Party Computations for Privacy-Preserving Machine Learning," *Procedia Computer Science* 213 (2022), https://www.sciencedirect.com/science/article/pii/S1877050922017914.

	unauthorised access.	Implemented by Apple.[113]	
Zero-knowledge Proofs	Involve one party confirming the authenticity of a piece of content to another party without revealing any other information.	Strong privacy guarantees. Enables verification without revealing content. Suitable for E2EE environments.	Complex implementation. Computationally intensive. Limited to predefined tasks.

5.2 Encryption-based Moderation Solutions

Encryption-based techniques use advanced cryptographic algorithms to enable detection tools to assess whether flagged content could be illegal, while preserving user data from unauthorised access and ensuring E2EE environments remain uncompromised.

5.2.1 Homomorphic Encryption

Homomorphic Encryption (HE) enables computations to be performed on encrypted data without decryption. For moderation purposes, this means media can be analysed for harmful content in a way that is compatible with E2EE protocols. HE can be implemented in different forms, including:

- **Fully Homomorphic Encryption (FHE)**, which supports unlimited operations on encrypted data.
- **Somewhat Homomorphic Encryption (SHE)**, which allows for a limited number of operations.
- **Partially Homomorphic Encryption (PHE)**, which permits specific operations (such as addition or multiplication).[114]

[113] Apple Security Engineering and Architecture, "Security research on Private Cloud Compute", *Apple Security Research*, 24 October 2024, https://security.apple.com/blog/pcc-security-research/.
[114] IEEE Digital Privacy, "Types of Homomorphic Encryption", https://digitalprivacy.ieee.org/publications/topics/types-of-homomorphic-encryption.

A significant benefit of HE relates to how any detection systems implemented with these techniques will preserve the encrypted nature of the content throughout the entire screening process, meaning that no party can access the underlying data.[115] Unlike traditional moderation solutions – under which adversaries can study and exploit detection rules – HE adds a fundamental mathematical barrier to evasion, which is particularly valuable given the sophisticated tactics employed by malicious actors outlined in Section 2. Indeed, the Information Commissioner's Office has specifically acknowledged HE as a valuable privacy-enhancing technology that can help organisations comply with data protection requirements while fulfilling their legal safety obligations.[116]

Nevertheless, HE operations can be thousands of times slower than non-cryptographic methods, making them less suitable for real-time analysis.[117] At the same time, most practical HE implementations support only basic content matching, with the more sophisticated contextual analysis required for effective moderation still beyond the capabilities of current proposals.[118]

5.2.2 Message Franking

Message Franking (MF) enables the verifiable reporting of harmful content in encrypted communications by cryptographically linking messages to their senders while preserving overall confidentiality. This approach addresses the challenge of abuse reporting in E2EE platforms without compromising legitimate communications.

While Facebook Messenger has already introduced this scheme, researchers have also integrated MF with Searchable Symmetric Encryption methods (see Section 5.2.4) for both proactive and reactive moderation.[119] This latter proposal uses designated verifier signatures to control who can check the information, while Signatures of Knowledge (SoKs) are included to strike a balance between holding users responsible and protecting their privacy. Finally, it has a smart filtering method to block fake reports. A layered

[115] Zheng et al., (2023), 5785; Saransh Gupta, Rosario Cammarota and Tajana Šimunić, "MemFHE: End-to-end Computing with Fully Homomorphic Encryption in Memory," *ACM Transactions on Embedded Computing Systems* 23, No. 2 (March 2024), https://dl.acm.org/doi/10.1145/3569955.
[116] ICO (2023), 30.
[117] Internet Society, "Homomorphic Encryption: What Is It, and Why Does It Matter?", 9 March 2023, https://www.internetsociety.org/resources/doc/2023/homomorphic-encryption/.
[118] Ibid.
[119] Peng Jiang, Baoqi Qiu and Liehuang Zhu, "Report When Malicious: Deniable and Accountable Searchable Message-Moderation System," *IEEE Transactions on Information Forensics and Security* 17 (2022), 1598, https://ieeexplore.ieee.org/document/9758811.

implementation of MF with ZKPs (see Section 5.2.6) could help achieve more robust privacy-preserving moderation.[120]

One advantage of MF is that it enables users to report abusive messages with cryptographic proof that the message was received, allowing platforms to verify reports and take appropriate action.[121] MF also ensures that regular communications remain private despite enabling verification of reported abuse, while providing cryptographic assurance that the reported content was sent.[122]

However, MF relies on users reporting harmful content after exposure rather than preventing exposure entirely. This limitation is particularly significant for content such as CSAM, where vulnerable children may be unable to report such activities. MF is also most effective for content types that users can easily recognise as harmful, such as text and static images, as opposed to more complex content types (e.g. encrypted video streams).

5.2.3 Private Set Intersection

PSI allows two parties to find the intersection of their datasets without revealing non-matching elements. When applied with detection tools, PSI-based methods enable platforms to detect whether user content matches a database of known harmful material without sharing either the full content or the database.[123]

One notable application of PSI was in Apple's proposed CSAM detection system, which set a threshold to decide when content needed to be flagged. PSI helped ensure that only shared, relevant data was identified, while protecting users' private information.[124] Although this scheme was never released, it demonstrated the technical feasibility of PSI for privacy-preserving content moderation.

Some PSI protocols can be computationally and memory-intensive, especially for large datasets or strong security models.[125] Optimisations reduce these overheads, but may lead

[120] CETaS focus group, 4 March 2025.
[121] Matthew Gregoire, Margaret Pierce and Saba Eskandarian, *Onion Franking: Abuse Reports for Mix-Based Private Messaging* (December 2024), https://eprint.iacr.org/2024/1965.
[122] Ibid.
[123] Scheffler and Mayer (2023), 17-18; Hawkes et al., (2024), 3.
[124] Abhishek Bhowmick et al., *The Apple PSI System* (July 2021), https://www.apple.com/child-safety/pdf/Apple_PSI_System_Security_Protocol_and_Analysis.pdf.
[125] Moni Naor and Benny Pinkas, "Oblivious transfer and polynomial evaluation," *STOC '99: Proceedings of the thirty-first annual ACM symposium on Theory of Computing* (May 1999), 246, https://dl.acm.org/doi/pdf/10.1145/301250.301312.

to trade-offs between security and efficiency.[126] Like all hash-based approaches, PSI can only be used with tools that detect previously identified harmful content – which cannot be used to address the distribution of novel material.

However, the strengths of PSI lie in its ability to facilitate hash matching without exposing either the hash database or non-matching content.[127] This protection is particularly important for highly sensitive repositories such as those containing CSAM fingerprints, where minimising access protects detection integrity and prevents unauthorised use. Participants in our focus group identified PSI as particularly promising, due to these unique advantages.[128]

Therefore, E2EE platforms should prioritise efforts to test content detection tools with PSI techniques in controlled environments before considering the wider implementation of their services, due to their strong privacy guarantees. In particular, services could explore the effectiveness of these methods in assisting pre-upload content screening of illegal material.

5.2.4 Searchable Symmetric Encryption

Searchable Symmetric Encryption (SSE) enables users to search for keywords in encrypted text without decrypting it. For content moderation purposes, this allows platforms to identify potentially harmful messages in encrypted communications without accessing the full content.[129]

The SSE solutions that have been proposed involve a combination of these techniques and asymmetric MF (see Section 5.2.2). This would involve special designated verifier signatures and encryption to allow keyword-based searches. It would include features to block fake messages and would use SoKs to balance the accountability and privacy of users.[130]

In terms of advantages, SSE allows for fast searches and requires significantly fewer steps than searching through all the data in question, thereby making it more efficient.

---

[126] Daniel Morales, Isaac Agudo and Javier Lopez, "Private set intersection: A systematic literature review," *Computer Science Review* 49 (August 2023), 15-17, https://doi.org/10.1016/j.cosrev.2023.100567.
[127] Hawkes et al., (2024), 10.
[128] CETaS focus group, 4 March 2025.
[129] Licheng Ji et al., "Verifiable Searchable Symmetric Encryption Over Additive Homomorphism," *IEEE Transactions on Information Forensics and Security* 20 (January 2025), 1320-1321, https://ieeexplore.ieee.org/document/10827839.
[130] Jiang et al, (2022).

Furthermore, SSE enables verifiable reporting of harmful messages without compromising E2EE protocols and prevents false reporting through cryptographic verification.

Despite these benefits, the focus on keyword matching means that SSE often struggles with contextual understanding – making it less effective in nuanced moderation processes or in reviewing images and videos. Metadata leakage is also a risk, with unauthorised users potentially able to observe search and access patterns (e.g. the ways in which systems or moderators retrieve and interact with content) and thereby improve their evasion tactics.

5.2.5 Secure Multi-party Computation

Secure Multi-party Computation (SMPC) allows multiple parties to jointly analyse content without exposing private data. Despite tasks being shared across the process, SMPC does not permit those involved to learn anything beyond the final outcome of the process (e.g. the content does not match an illegal database) due to the use of encryption methods.[131]

SMPC can incorporate techniques such as garbled circuits to provide even greater privacy guarantees by allowing computations to happen without revealing private inputs, while secret sharing splits data so that no single party has access to the full information.[132] It can also support diverse content moderation requirements – including text analysis, image classification and multimodal content assessment.[133] Finally, by distributing computation and data across multiple parties, SMPC reduces reliance on any single trusted entity. This applies equally to contexts in which security can be maintained even if a subset of participants is compromised.[134]

In terms of drawbacks, SMPC protocols typically require multiple rounds of interactive communication between participating parties, creating significant overheads and making it impractical in distributed environments.[135] SMPC operations are also significantly more computationally expensive than non-cryptographic methods, adding performance bottlenecks to complex moderation tasks.[136] As with many other techniques listed in this

[131] Zapechnikov (2022).
[132] Ibid, 525-526.
[133] Ibid, 524.
[134] Chuan Zhao et al., "Secure Multi-Party Computation: Theory, practice and applications," *Information Sciences* 476 (February 2019), 358, https://www.sciencedirect.com/science/article/pii/S0020025518308338.
[135] Alex Haynes, "Multi-Party Computation: A Double-Edged Sword for Cybersecurity", *United States Cybersecurity Magazine* 13, No. 42 (2024), https://www.uscybersecurity.net/csmag/multi-party-computation-a-double-edged-sword-for-cybersecurity/.
[136] Ibid.

section, SMPC lacks standardised protocols when used for moderation purposes, which can undermine their efficiency when implemented on different types of networks.

Subsequently, ISO/IEC JTC 1/SC 27 – which is responsible for the standardisation of online cryptographic privacy protections – should develop new standards of protocols and interfaces for SMPC and other privacy-enhancing technologies applied within content moderation contexts.[137] This would help ensure consistency, interoperability and scalability across platforms.

5.2.6 Zero-knowledge Proofs

When combined with detection tools, ZKPs allow one party to verify to another that the scanned content does or does not match known harmful material, without revealing the content or the database of harmful material it is compared against.[138]

Researchers have proposed a ZKP-based moderation process whereby a hash database cryptographically stores content, using a secret key to keep it secure.[139] To prove that the content is not matched in the database, the protocol generates a cryptographic proof of such confirmation without disclosing additional details on the content analysed.

ZKPs provide cryptographic certainty that only the necessary information is revealed during verification processes.[140] They can also improve the privacy properties of perceptual hash matching by allowing verification without exposing the hash database. Finally, recent advances have made ZKP systems increasingly practical for real-world applications. Recent tests on a proposal by Succinct Labs has shown proof generation taking only 147 milliseconds and verification taking just 66 milliseconds.[141]

While more efficient than some alternatives, ZKPs still impose computational overheads. Further work is needed to improve non-interactive ZKP efficiency and standardise implementation protocols. Despite these limitations, ZKPs were another approach our focus group participants mentioned as having the potential to overcome many of the thorny moderation challenges in E2EE spaces.[142] This was due to their ability to prevent the parties

[137] ISO, "ISO/IEC JTC 1/SC 27: Information security, cybersecurity and privacy protection," https://www.iso.org/committee/45306.html.
[138] ICO (2023), 34-35; Bartusek et al., (2023), 6.
[139] Scheffler, Kulshrestha and Mayer (2023), 264.
[140] Ibid, 258.
[141] Scheffler, Kulshrestha and Mayer (2023).
[142] CETaS focus group, 4 March 2025.

involved from learning anything other than that the content in question did not match an illegal database, reducing the risk that detection technologies will be misused to gather unintended information about users or the content they are sharing.

Accordingly, E2EE platforms should test detection tools in tandem with ZKPs in controlled environments before considering wider implementation on their services. As with PSI-based methods, services could explore the effectiveness of ZKPs in assisting pre-upload content screening processes.

5.3 Other Promising Moderation Solutions

5.3.1 AI image-to-text moderation

Image-to-text moderation uses AI and ML techniques to convert images into text formats. Some tech companies – such as Amazon and Microsoft – have already implemented partial methods of this kind. These detection systems scan images for any text contained within them, before checking for inappropriate language, hate speech or other violations of platform policies against databases of banned words.[143]

From a moderator perspective, this process could help reduce the risk of exposing a human in the loop to potentially harmful images that could cause psychological damage. Furthermore, the text component would be more sharable than the imagery itself, opening up new options for sharing datasets for ML purposes.

However, while exposing a text description of an image would likely be less of a privacy intrusion than sharing the image itself, these types of solutions have yet to be tested in E2EE networks and require further research to understand their impact on such protocols. They are also limited to image formats, undermining their wider applicability for moderation tasks.

5.3.2 Anonymous blocklisting

Anonymous blocklisting is a privacy-preserving method for sender-anonymous messaging, allowing recipients to block unwanted senders without revealing their identities. It is particularly useful in E2EE environments, where conventional blocking methods could compromise anonymity. The challenge lies in balancing sender privacy with abuse

[143] AWS, "Amazon Rekognition Content Moderation", https://aws.amazon.com/rekognition/content-moderation/; Microsoft Learn, "Learn image moderation concepts", 28 August 2024, https://learn.microsoft.com/en-us/azure/ai-services/content-moderator/image-moderation-api.

mitigation, as traditional moderation relies on user identification – which conflicts with the principles of encrypted communication.

One proposal uses cryptographic techniques such as group signatures to allow users to remain anonymous while proving their identity within a group. Verifier-local revocation then prevents the exposure of metadata if someone loses their rights or access, keeping their information private.[144] This method can be used in content moderation by filtering messages from blocked senders while preserving anonymity. Platforms can also enforce blocking without identifying the users involved. Cryptographic verification ensures privacy while allowing moderation, and the system is scalable, enabling deployment across large networks.[145]

Despite its advantages, anonymous blocklisting introduces computational overheads, as cryptographic verification increases processing demands. For systems that rely heavily on metadata to identify harmful behaviour (e.g. E2EE services), service providers could be severely limited in their ability to identify malicious actors. Such individuals may also attempt to bypass blocking by creating new accounts, necessitating additional protections that incur costs. Future refinements are also needed to address platform-side denial-of-service risks and message attribution issues.[146]

5.3.3 Federated Learning

Federated Learning (FL) allows ML models to be trained across multiple devices or servers holding local data samples without exchanging the data itself. In content moderation, FL enables collaborating operators to co-train models while keeping sensitive user data on their own devices.[147]

Proposals for privacy-preserving FL designs have involved the integration of Central Differential Privacy methods for harmful content detection.[148] This system allows users to control the development of personalised local detection models based on their own dataset

[144] Nirvan Tyagi et al., "Orca: Blocklisting in Sender-Anonymous Messaging," *Proceedings of the 31st USENIX Security Symposium* (2022), https://www.usenix.org/conference/usenixsecurity22/presentation/tyagi.
[145] Ibid, 2,303.
[146] Ibid, 2,302-2,303.
[147] Leonidou et al., (2023).
[148] Mohammad Naseri, Jamie Hayes and Emiliano De Cristofaro, "Local and Central Differential Privacy for Robustness and Privacy in Federated Learning," *Network and Distributed Systems Security Symposium* (April 2022), https://www.ndss-symposium.org/wp-content/uploads/2022-54-paper.pdf.

labels, while still contributing to a centralised moderation model and preserving user privacy.

FL aligns with data minimisation principles by keeping user data on local devices, reducing the need for centralised data collection and storage. This approach enables platforms to develop and improve moderation without requiring users to upload sensitive content to central servers, addressing a fundamental privacy concern. By incorporating differential privacy techniques, FL can also provide formal privacy guarantees against common attack vectors such as membership inference. [149]

Despite these advantages, FL systems can be susceptible to data poisoning attacks, in which malicious users deliberately provide misleading training examples to compromise the centralised model.[150] As with all community moderation approaches, FL relies on users flagging harmful content to define the type of data from which the system learns. [151] When different users define harms in their own way, this can create messy and unreliable data. Such inconsistencies make it harder for the centralised model to combine information properly, weakening its ability to moderate content effectively.

5.3.4 Trusted Execution Environments

Trusted Execution Environments (TEEs) offer a promising compromise between cloud-based processing capabilities and privacy preservation. TEEs provide an isolated computational space in which sensitive operations can be performed with hardware-level protection against unauthorised access – even from the system operators themselves.[152] One example of this approach is the Apple Private Cloud Compute, which is designed to provide secure and private processing of AI models in the cloud.[153]

TEEs could enable online platforms to run sophisticated detection algorithms in the cloud while maintaining strong privacy guarantees. This is because content would only be decrypted within the secure enclave, with access to unencrypted content restricted to cases in which there are positive matches with harmful databases.[154] TEEs also facilitate the use of complex moderation techniques that might be impractical for on-device implementation due to resource constraints.[155] Finally, new AI models could be set to

[149] Ibid, 14-15.
[150] Naseri, Hayes and De Cristofaro (2022), 12-13.
[151] CETaS focus group, 4 March 2025.
[152] ICO (2023), 36-38.
[153] Apple Security Engineering and Architecture (2024).
[154] Scheffler and Mayer (2023), 18.
[155] ICO (2023), 36-37.

generate large volumes of content in a secure environment, with the content being assessed by one or more AI classifiers to check if harmful material was created – all without sharing or exposing anyone to such content. This would help reduce the risk of the model subsequently being misused once released.

Despite their promise, TEEs face important limitations. For example, they require specialised hardware support, creating deployment challenges across heterogeneous device ecosystems.[156] Additionally, various side-channel vulnerabilities in TEE implementations have been identified, potentially compromising privacy guarantees through sophisticated attacks.[157]

[156] Scheffler and Mayer (2023), 18.
[157] Ibid.

6. Conclusion

As debates persist over how to achieve an optimal balance in user privacy and safety when moderating illegal content on social media and E2EE platforms, there is an urgent need for innovation to address increases in real-world harms tied to the proliferation of such material.

This report provides a series of policies, frameworks and tools that tech companies can implement across a variety of online domains to better protect users while minimising privacy intrusiveness.

Regardless of which approaches different platforms implement, there will always be a need for iteration – as innovation progresses and malicious actors find new ways to circumvent existing techniques. For this reason, it is essential for both tech platforms and those who hold them accountable to frequently reflect on moderation practices. Indeed, the various solutions detailed in this report should not be perceived as a panacea for the problem of illegal online content but serve as a foundation for further experimentation.

Throughout this project, the study team identified several gaps for future research and testing. These include:

- Exploring the potential benefits of enhancing AI-based moderation systems with contestability algorithms, which are designed to improve the decision-making process behind content removal.[158]
- Optimising the efficiency of HE operations through improved algorithms and hardware acceleration.
- Testing SSE-based moderation systems integrated with ML techniques to enhance their contextual understanding capabilities.
- Developing non-interactive ZKP protocols that reduce vulnerability to timing attacks and more efficient implementations that address current performance limitations.

We urge the UK Government and platform providers to carefully consider the recommendations in this report, to help keep people safe online now and in the future.

[158] Elkin-Koren (2020).

About the Authors

Sam Stockwell is a Research Associate at the Centre for Emerging Technology and Security (CETaS). His research interests focus on how AI is affecting the online information ecosystem, particularly in relation to misinformation and disinformation, the moderation of harmful content and election interference.

Georgia Wake is a Senior Analyst within the UK Government. Her research interests focus primarily on the exploitation of platforms and emerging technologies for the purposes of terrorist recruitment and radicalisation.

Dr Tooska Dargahi is a Senior Lecturer in Cyber Security at Manchester Metropolitan University (MMU). Her research interests primarily focus on user-centric privacy and security across diverse interconnected environments, including the Internet of Things, smart cities, healthcare and cyber-physical systems. She actively contributes to public engagement activities aimed at raising awareness of the security and privacy challenges faced by users in modern digital society.

Dr Oluwaseun Ajao is a Senior Lecturer in Data Science & AI at MMU. His research interests include misinformation, natural language processing and artificial intelligence. He is part of the Taste of the Algorithm content moderation project, funded by the Centre for Advanced Internet Studies Germany. He is on the editorial boards of the *Journal of Artificial Intelligence and Robotics*, *AI Insights* and the *International Journal of Information and Communication Sciences*.

Dr Annabel Latham is a Senior Lecturer in Computer Science and the Information Systems Curriculum Leader at the School of Computing, Mathematics & Digital Technology at MMU. Annabel's research interests include conversational agents, intelligent tutoring systems, AI in education and public trust in AI systems. She is on the editorial board of *IEEE Transactions on AI*.

Ahmed Danladi Abdullahi is a PhD candidate in the Department of Computing and Mathematics at MMU.

Dan Sexton is the Chief Technology Officer at the Internet Watch Foundation (IWF). He is responsible for Information Technology, Cybersecurity and Software Development, and he leads the IWF's software engineering work.

CETaS

Centre for Emerging Technology and Security

RESEARCH REPORT