Annex 4: Additional policy considerations for specific measures

Additional Safety Measures

Statement Annex

Published 16 November 2026

For more information on this publication, please visit ofcom.org.uk

Contents

A1. Proactive technology: CSAM proactive technology technical annex 3

A2. Proactive technology: draft guidance changes 27

A3. Human review for proactive technology 34

A4. Hash matching for terrorism content: detailed reasoning on effectiveness 46

A5. Assessment of Tech Against Terrorism's Hash List API service 53

A1. Proactive technology: CSAM proactive technology technical annex

Warning: this annex contains content that may be upsetting or distressing.

Introduction

A1.1 This annex sets out our reasoning for why we have recommended certain classes and subclasses of technology in Measure ICU PT1: Deployment of proactive technology to detect image-based child sexual abuse material (CSAM). Firstly, we set out our approach to assessing classes of technology and our understanding of the harm the technology would be detecting. Secondly, we set out our assessments of each class/subclass we have concluded are capable in principle of being accurate, effective and lacking in bias where they meet the proactive technology criteria in regard to their ability to detect unknown image-based CSAM.[1] We refer to these classes and/or subclasses as 'specified technologies'.

A1.2 Table A1.1 contains a definition and summary of reasoning for the specified technologies we are recommending for the purpose of detecting unknown image-based CSAM.

[1] We define unknown image-based CSAM as content that is not included in any reference database of hash values for CSAM used by the provider in order to comply with ICU C9 of the Illegal content Codes of Practice for user-to-user services. We acknowledge that proactive technology that is designed and deployed to detect unknown CSAM is also capable of detecting known CSAM. However, to ensure the detection and removal of CSAM is as quick and effective as possible, we consider it is proportionate that in-scope services continue to use our recommended hash matching measure for known CSAM (ICU C9) alongside proactive technology for unknown CSAM. Hash matching is a well-established industry practice that is highly accurate at detecting known CSAM and enables automated removal and reporting of that content without further layers of moderation or human review. Hashmatching is also necessary to protect child users who may appear to be older than they are, as review of the content may not be sufficient on its own to identify it as CSAM. Proactive technology is vital for the detection of unknown CSAM but should be implemented alongside hash matching as it may not always detect known content with the same level of accuracy or confidence as hash matching.

Table A1.1: Summary of specified technologies and reasoning for inclusion

Specified technologies	Definition[2]	Summary of reasoning
Text Matching	'Text Matching' represents one subclass of the 'Rules-Based and Heuristic Technologies' technology class (see the section entitled 'Scope of the assessment'). Exact matching, approximate matching, or pattern matching technologies based on rule-based or heuristic matching approaches to identify predefined textual elements and patterns within content, including words, phrases, character sequences, uniform resource locators (URLs), and domain names.	We have found that textual indicators associated with image-based CSAM are sufficiently associated with the target harm to support meaningful detection outcomes and that such indicators can be operationalised through appropriately configured text matching systems. We have also found that identified limitations may be mitigated through the use of appropriate indicator datasets, ongoing maintenance and deployment practices consistent with the proactive technology criteria. Our analysis shows that a number of high quality and well-maintained datasets containing these textual indicators are available. Taking these factors into account, we conclude that text matching technologies are capable in principle of being sufficiently accurate, effective and lacking in bias for the purpose of detecting unknown image-based CSAM where they meet the proactive technology criteria.

[2] These definitions were developed internally for the purposes of classifying and assessing proactive technologies as part of our technical analysis and do not therefore derive from a single external source.

Specified technologies	Definition[2]	Summary of reasoning
Machine learning	'Application-specific Machine Learning' is a technology class (see the section entitled 'Scope of the assessment'). Technologies based on models developed or adapted using training data to identify, classify, or assess features, attributes, or characteristics of target content, and whose intended tasks, decision criteria, and expected outputs are established through model development or adaptation.	Application-specific machine learning technologies are capable of directly analysing image and video content and identifying visual features associated with image-based CSAM. Where representative training data is available, such technologies are capable of achieving high levels of classification performance across image-classification tasks. While access to such datasets can present challenges in the case of CSAM, specialist organisations such as NCMEC have facilitated access to representative CSAM datasets and related training resources. We also consider that the limitations associated with these approaches can be mitigated through appropriate dataset governance, calibration, testing, monitoring and maintenance practices. Taking these factors together, we conclude that application-specific machine learning technologies are capable in principle of accurately and effectively supporting detection of image-based CSAM, in a manner that lacks undue bias, where they meet the proactive technology criteria.

A1.3 We are doing further work on assessing other classes/subclasses of technologies, including large language models and multimodal foundational models, and we may consult on adding other specified technologies to this measure in the future.

Our approach to technical assessments

Purpose

A1.4 The Act requires Ofcom to have regard to the degree of accuracy, effectiveness and lack of bias achieved by the proactive technology in question in deciding whether to recommend it in a Code of Practice.[3] 'Accuracy', 'effectiveness' and 'lack of bias' are not defined in the Act. We take each to have their ordinary meanings in English:

- 'Accuracy' is to do with 'being exact or correct'.
- 'Effectiveness' is to do with being 'successful in producing a desired or intended result'.

[3] Schedule 4 to the Act, paragraph 2(c).

- ‘Lack of bias’ is to do with being equitable and fair.

A1.5 This annex sets out how we have had regard to the degree of accuracy, effectiveness and lack of bias of proactive technology in order to recommend it for the purpose of detecting unknown image-based CSAM in ICU PT 1.

A1.6 We considered the use of metrics in assessing the classes/subclasses of technologies, however we have not taken this approach for the following reasons.

A1.7 Firstly, proactive technology can be accurate without being effective and vice versa. Proactive technology that is not effective at detecting relevant content in a real-world context may still be considered ‘accurate’ under commonly used accuracy metrics if a harm has low prevalence on a service. As an illustrative example, consider a dataset of 100 items with five illegal data points and 95 benign ones:

a) ‘Technology A’ correctly identifies 90 out of 100 of the total data points, achieving a 90% accuracy rate. All five instances of illegal content are correctly identified, but there are additional false positives included in the outputs.

b) ‘Technology B’ labels all 100 data points in the dataset as benign. It achieves a 95% accuracy rate but fails to identify any of the illegal content.

A1.8 Therefore, accuracy must be considered in combination with effectiveness. Setting thresholds for accuracy without thresholds for effectiveness could incentivise providers to prioritise reducing false positives over identifying harmful content, thereby leaving users less safe. However, it is not possible to set thresholds for effectiveness that apply across all services. This is because to measure how effective proactive technology is at detecting image-based CSAM, as a proportion of all other types of content, we would need a robust estimate of the total volume of image-based CSAM. We are not aware of any robust, industry-accepted metric on the prevalence of image-based CSAM on user-to-user services. In any event, it appears likely that the prevalence varies unpredictably from service to service.[4] Even if we could show prevalence across the sector, it follows that it may not be appropriate to use this to weigh up accuracy and effectiveness for a particular service.

A1.9 Secondly, even if we specified theoretical target thresholds for accuracy and effectiveness, the same technology may generate different accuracy and effectiveness results depending on the content against which it is evaluated.

a) Where prevalence of image-based CSAM content is low, a random sample of content taken from a given service may not contain any, and would therefore provide limited insight into the real-world performance of proactive technology.

b) The measured accuracy and effectiveness of a technology will depend on the content of the dataset against which it is evaluated. As a result, the same technology may generate different results when evaluated on different datasets. Thresholds specified by Ofcom would therefore only provide a meaningful basis for comparison if technologies were evaluated under comparable conditions, such as the use of the same test dataset.

c) The nature and prevalence of harm, as well as the volume of non-relevant content that may be mistakenly identified as relevant content, are likely to differ across services. As a result, technologies that perform well under a standardised test and meet prescribed thresholds for accuracy and effectiveness may not be appropriate or effective when

[4] See, for example, paragraph A4.47(i) of Annex 5 of our December 2024 Statement on Protecting People from Illegal Harms Online (December 2024 Statement).

deployed on a particular service. Consequently, we would not expect universal thresholds based on standardised test conditions to provide a reliable indication of performance in practice.

A1.10 Even if technology met thresholds for accuracy and effectiveness when we assessed it, it may not meet these thresholds when applied on services due to the above contextual changes.

A1.11 Lastly, we have estimated that the number of online services likely to be in scope of the Act could be over 100,000. While we apply our code measures by size and/or risk, we consider this scope to be too broad for the purpose of setting thresholds. We do not consider it would be proportionate or effective for Ofcom to assess technology against different thresholds for every combination of: service type; types of content; different deployment approaches and different user bases on services (their cultures, interests, languages and vulnerabilities). These factors vary from service to service and evolve regularly, meaning any potential threshold we assessed technology against would quickly become out-of-date.

A1.12 Overall, these considerations arise even when assessing individual technologies or products, but become more pronounced when assessing at the level of a technology class, which may encompass a range of approaches, implementations and deployment contexts.

A1.13 Considering we are unable to use metrics to assess the degree of accuracy, effectiveness and lack of bias of technology, we have taken an alternative approach. The following sub-sections outline this approach in detail.

Analytical framework for the assessment

A1.14 Having regard to these considerations, we have conducted a technical assessment of whether, in principle, certain classes of proactive technology are capable of being sufficiently accurate, effective and lacking in bias (as defined above) for the purpose of detecting unknown image-based CSAM where they meet the criteria as set out in chapter 9: Deployment of proactive technology to detect image-based CSAM. We have also considered the availability of those technologies in our assessment.

A1.15 As explained in chapter 9: Deployment of proactive technology to detect image-based CSAM, we consider that providers ensuring specified technologies meet the proactive technology criteria is important to ensure that they are accurate, effective and lacking in bias in the context of actual service deployment. As explained in the chapter, we have developed the proactive technology criteria drawing on evidence from national and international frameworks on responsible artificial intelligence (AI) and automated content moderation, and have considered common challenges in the design and deployment of technologies in devising the criteria.

A1.16 To support our technical analysis, we have developed an analytical framework which groups technologies into technology classes and, where appropriate, subclasses. The analytical framework has been developed to support regulatory capability assessment and is not intended to provide a technically exhaustive taxonomy of proactive technologies, machine-learning techniques or underlying algorithmic approaches.

A1.17 In developing the technology classes, we have considered the primary signal analysed by a technology, the primary moderation or detection function it performs, and the factors most relevant to the assessment of its accuracy, effectiveness and lack of bias. This has enabled us to assess technologies with similar operational characteristics, strengths,

limitations and factors affecting their accuracy, effectiveness and lack of bias on a consistent basis.

A1.18 Where technologies within a technology class exhibit materially different capabilities, limitations or deployment considerations relevant to the assessment, we further distinguish between subclasses in order to assess those differences separately.

Scope of the assessment

A1.19 The broader assessment comprises the following technology classes and subclasses:

- **Rules-Based and Heuristic:** technologies which encompass techniques that identify, classify or analyse content using predefined rules, patterns, matching criteria, thresholds or decision procedures. These approaches determine outcomes by applying explicitly specified detection logic to content or associated signals. This includes the subclasses of:
  - Text Matching
  - Perceptual Hashing[5]
  - Cryptographic Hashing[6]
  - Rule Engines[7]
- **Application-specific Machine Learning**: technologies based on models developed or adapted using training data to identify, classify, or assess features, attributes or characteristics of target content, and whose intended tasks, decision criteria and expected outputs are established through model development or adaptation.
- **Foundation Models**: technologies based on machine learning models trained on broad and diverse datasets and developed for general-purpose use across a range of downstream tasks. The tasks performed, decision-making logic or outputs generated by such models are specified, modified or constrained through prompting, fine-tuning, deployment configuration or other adaptation techniques. This includes the subclasses of:
  - Large Language Models (LLMs); and
  - Multimodal Foundation Models (MFMs).
- **Behavioural and Network-Based**: technologies which encompass techniques that analyse behavioural, relational and network signals to identify suspicious, anomalous or coordinated activity. They operate by examining how users, accounts, devices or networks behave and interact, rather than analysing the content itself.

A1.20 We have also considered the content modality to which a technology was applied (for example, image, video, text, audio or multimodal).

A1.21 The capability of a technology to achieve sufficient levels of accuracy, effectiveness and lack of bias may vary depending on the modality being assessed. Accordingly, our assessment considers technologies in the context of the specific harm manifestation and

[5] Determines whether a given file is perceptually similar to a hashed file by setting a threshold that is used to determine when there is sufficient similarity between the hashes to be considered a (near) match. This allows modifications from an original file to be detected.

[6] Only detects files that are an exact match. This means that perpetrators may slightly alter images to circumvent it.

[7] Rule engines: technologies that evaluate explicitly defined rules against one or more inputs and execute predetermined actions or produce decisions when specified conditions are satisfied.

content modality relevant to the measure, rather than assuming that conclusions reached for one modality would necessarily apply to another.

Evidence base

A1.22 In carrying out our assessment, we drew on a range of evidence sources, including academic literature, evidence and intelligence published by law enforcement and civil society organisations, information provided by regulated services and technology providers, responses to our June 2025 Additional Safety Measures Consultation (June 2025 Consultation), Ofcom-commissioned research, and other publicly available sources.

Considering accuracy, effectiveness and lack of bias

A1.23 For each technology class, we have examined the available evidence relating to inherent strengths and limitations, the risks of false positive and false negative detections, and information about deployment and operational use of approaches within each technology class. We have considered these factors because they provide evidence of how a technology functions, the conditions under which it performs well or poorly, and the sources of error or systematic differences that may affect performance outcomes. As such, they are central to an assessment of accuracy, effectiveness and lack of bias.

A1.24 In considering accuracy, we have evaluated the available evidence relating to the degree to which the core capabilities of approaches within each technology class support correct identification of the target harm, including the relevance of the signals we expect to be available to the technology and the potential for detection errors.

A1.25 In considering effectiveness, we have evaluated the available evidence relating to the degree to which the core capabilities of approaches within each technology class support the intended detection objective, taking account of their strengths, limitations and known constraints.

A1.26 In considering lack of bias, we have evaluated the available evidence relating to the degree to which, based on their core capabilities, approaches within each technology class appeared capable of operating without creating systematic differences in outcomes across relevant groups, languages, contexts or content types.

Operational and deployment considerations

A1.27 In addition to considering the core capabilities of each technology class, we have examined evidence relating to how such technologies are deployed and used in practice. This includes consideration of the availability of third-party solutions, existing deployment evidence, practical implementation requirements, operational constraints, and the extent to which technologies are typically deployed as part of wider moderation workflows. These considerations have informed our assessment of whether technologies appear capable of achieving a sufficient degree of accuracy, effectiveness and lack of bias when deployed in real-world settings.

Layering analysis

A1.28 Our analysis has also considered whether the available evidence indicated that approaches within a technology class were capable, in principle, of being sufficiently accurate, effective and lacking in bias in isolation where they meet the proactive technology criteria, or whether achieving those outcomes depended upon deployment alongside a complementary proactive technology class (i.e. layering).

Assessment outcomes

A1.29 Following our analysis, we have assigned each technology class or subclass (as relevant) one of three capability outcomes:

- **Capable in Principle**: based on our analysis of the available evidence, for this harm manifestation, we consider the technology class or subclass capable in principle of being sufficiently accurate, effective and lacking in bias for the purpose of detecting unknown image-based CSAM where it meets the proactive technology criteria. Although we identified limitations and risks, we did not consider these sufficient to prevent the technology from meeting the proactive technology criteria in principle.
- **Capable in Principle when Layered with one or more additional capable technology classes and/or subclasses**: based on our analysis of the available evidence, for this harm manifestation, we do not consider the technology class or subclass in isolation capable in principle of being accurate, effective and lacking in bias. However, where deployed alongside one or more additional capable proactive technology classes or subclasses, we consider the identified limitations may be sufficiently mitigated. Accordingly, we have considered the technology capable in principle of being sufficiently accurate, effective and lacking in bias where it meets the proactive technology criteria as part of a layered approach.
- **Not Recommended:** based on our analysis of the available evidence, for this harm manifestation, we do not currently consider this technology class or subclass is capable in principle of being sufficiently accurate, effective and lacking in bias even where it meets the proactive technology criteria. We do not have sufficient evidence to support that they are capable and/or identified significant limitations which we consider cannot be adequately mitigated through configuration choices, operational safeguards or layering with complementary technologies. We are therefore not recommending use of this technology class or subclass in our Codes (though service providers remain free to use the technology in question as a matter of their own commercial choice).

A1.30 These outcomes reflect our overall assessment, based on the available evidence, of the degree of accuracy, effectiveness and lack of bias that approaches within a technology class or subclass appear capable of achieving in principle where they meet the proactive technology criteria.

A1.31 Table A1.2 summarises each class and/or subclass of technology we have assessed and our conclusions.

Table A1.2: Summary of classes and/or subclasses of technology assessed and our conclusions.

	Application-specific Machine Learning	Foundation Models		Behavioural and Network Based	Rules Based and Heuristic			
		LLM	MFM		Cryptographic hashing	Perceptual hashing	Text matching	Rule engines
Unknown Image-based CSAM (Image)	Capable in principle	Further work needed	Further work needed	Capable in Principle when layered with one or more additional capable technology classes and/or subclasses	Not recommended	Not recommended	Capable in principle	Capable in principle when layered with one or more additional capable technology classes and/or subclasses
Unknown Image-based CSAM (Video)	Capable in principle	Further work needed	Further work needed	Capable in Principle when layered with one or more additional capable technology classes and/or subclasses	Not recommended	Not recommended	Capable in principle	Capable in principle when layered with one or more additional capable technology classes and/or subclasses

How image-based CSAM manifests and how proactive technology can identify this

A1.32 In this section we explain what image-based CSAM is and how it manifests online.

What is the harm

A1.33 Online image-based CSAM content is content amounting to an offence specified in any of paragraphs 2 to 4, 7 or 10 of Schedule 6 to the Act, so far as the risk in relation to those offences relates to CSAM in the form of photographs, videos or visual images.[8] It includes material depicting sexual activity involving children, or indecent or prohibited imagery of children, and can take the form of photographic images and videos, as well as non-photographic material such as drawings and animations.[9] [10] Image-based CSAM can also include AI-generated content.

A1.34 For the purposes of our assessment, unknown image-based CSAM refers to image-based CSAM that is not included in any reference database of hash values for CSAM used by the provider in order to comply with ICU C9 of the Illegal content Codes of Practice for user-to-user services (Illegal Content Codes). This includes genuinely novel CSAM content that has not been detected before, derivative content, transformed content, synthetic content, or content that is otherwise unsuitable for effective identification using perceptual hash matching techniques.

A1.35 While all image-based CSAM is illegal and can cause significant harm to victims and those that view it, unknown CSAM that is novel content is likely to be particularly high risk. Such content may depict children that are currently experiencing or have recently experienced harm. Detecting such content may therefore enable current victims of CSA to be safeguarded and prevent ongoing harm.

How the harm manifests online

A1.36 Image-based CSAM manifests through a range of image and video-based formats. This may include genuinely novel abuse imagery depicting previously unidentified acts of victimisation, derivative content such as screenshots, frame grabs, clips, thumbnails and other artefacts generated from known image-based CSAM (including where such content may be present within or derived directly from a livestream),[11] [12] transformed content

[8] These are the offences within Schedule 6 to the Act which are, by definition, about images. We have not included the offences in paragraphs 1 (obscene publications) or 8 (paedophile manuals) even though those offences are included in our definition of the harm for risk assessment purposes. This is because although it is technically possible for these offences to be committed through photographs, videos or other visual images the technology we have considered in this schedule is not necessarily well-adapted to detecting them. We do not consider it is likely that any service will be at risk of image-based CSAM based solely on its risk of image-based obscene publications or image-based paedophile manuals.

[9] Schedule 6 to the Act.

[10] For more information about the offenses relating to child sexual abuse material (CSAM) see: Ofcom, 2025. Protecting people from illegal harms online: Illegal Content Judgements Guidance (ICJG).

[11] WeProtect Global Alliance, 2026. Livestreaming.

[12] Internet Watch Foundation (IWF), 2025. 2025 Annual Data & Insights Report: Executive Summary, Severity of Images.

including edited or manipulated versions of known CSAM,[13] and fully synthetic CSAM content generated using AI.[14]

A1.37 Image-based CSAM will always include a depiction of a child (someone under 18 years old), whether real or synthetic. Visual features of individuals depicted in an image, including features associated with apparent age, can provide signals relevant to assessing whether the image may depict a child.

A1.38 As well as depicting a child, for an image to be CSAM it must also be indecent. For example, it may depict sexual activity, or children posing or acting sexually. CSAM will very often involve child nudity or partial nudity.[15] Other objects in an image may also help to indicate that an image could be CSAM. For example, a child on a bed, in a bedroom or bathroom (in combination with other indicators of indecency), or the presence of sex toys or certain bodily fluids.

A1.39 It may also be possible to detect CSAM images not by searching for the image itself, but for content linked to the image or referring to the image. This could include filenames, the labelling or title of content or comments posted on the content. For example, evidence suggests that offenders who share content and those who comment on it frequently discuss the age of the child or use coded or non-coded terms associated with CSAM.

A1.40 Offenders may also share URLs alongside CSAM which direct users to further CSAM. These URLs may be considered keywords associated with CSAM that could support the detection of a CSAM image.[16]

Specified technologies

A1.41 We have identified two specified technologies that we consider are capable in principle of being sufficiently accurate, effective and lacking in bias to for the purpose of detecting unknown image-based CSAM content where they meet the proactive technology criteria. In the following paragraphs, we explain what each of the recommended technologies are, and set out how they can detect each manifestation of image-based CSAM set out above.

Text matching for the purpose of detecting unknown image-based CSAM

A1.42 Text matching technologies focus specifically on the identification of textual elements and patterns and encompass techniques that detect predefined words, phrases, character sequences, alphanumeric identifiers, URLs, domain names and other structured textual

[13] IWF, 2025. 2025 Annual Data & Insights Report: Executive Summary, Analysis of Video Severity.
[14] IWF, 2025. 2025 Annual Data & Insights Report: Executive Summary, AI-generated Child Sexual Abuse Images: 2025 Data and Trends.
[15] We recognise that not all images involving child nudity are indecent.
[16] Ofcom's Illegal content Codes of Practice for user-to-user services already recommend the use of matching technology to detect and remove content matching listed CSAM URLs (Measure ICU C10).

elements. Matching may be exact or approximate and may involve regular expressions,[17] dictionaries, text normalisation[18] or other predefined matching criteria.

A1.43 Text matching is implemented in many major platforms and specialist child safety organisations that maintain and deploy rule-based, keyword-based or pattern-based moderation systems and resources intended to support content moderation, risk identification and policy enforcement.[19]

A1.44 In the context of image-based CSAM, text matching approaches may be applied to a wide range of textual elements associated with such content, including captions, comments, filenames, metadata, tags, usernames, profile information, URLs, accompanying communications and livestream chat.[20]

A1.45 In this respect, text matching approaches contribute to detection by identifying textual indicators associated with the likely presence of image-based CSAM rather than through direct analysis of image or video content itself.[21]

A1.46 The purpose of detecting textual indicators is to detect associated image or video content suspected to be image-based CSAM. Depending on the deployment context, the detected indicator may form part of the image or video content itself (for example through filenames, metadata or associated descriptive text) or may be used to identify corresponding image or video content contained within the same post, communication, message or livestream for subsequent assessment.

A1.47 Text matching encompasses a range of approaches that may be deployed individually or in combination within a moderation pipeline. These approaches vary in how they identify relevant indicators, but all operate by identifying predefined textual signals associated with a moderation objective. The principal approaches within the text matching subclass are:

- Exact matching, which identifies predefined keywords, phrases or character strings;
- Pattern matching, which seeks to identify indicators that follow a previously observed structure or format, even where the specific terms differ. For example, if a provider identifies that content associated with CSAM consistently uses terms following a particular combination of words, numbers or identifiers, pattern matching can be configured to detect additional instances that conform to that structure; and
- Similarity and approximate matching, which identifies indicators that are similar, but not identical, to predefined terms and may support detection of spelling variations, abbreviations, character substitutions or other forms of textual modification.

A1.48 When assessing the capability of the text matching subclass, we have considered the extent to which these approaches are capable of providing a sufficient degree of accuracy, effectiveness and lack of bias for the purpose of detecting unknown image-based CSAM, where they meet the proactive technology criteria.

---

[17] Regular expression (regex) refers to a text-matching technique that uses predefined patterns to identify words, phrases, or other character sequences. See: Luo, J. W., and Chong, J. J. R., 2020. Review of Natural Language Processing in Radiology, Neuroimaging Clinics of North America, 30 (4).

[18] Text normalisation is the process of converting text into a standardised format before analysis or matching, for example by making text lowercase, removing punctuation, or standardising alternative spellings.

[19] Discord, 2026. AutoMod FAQ; IWF, n.d. Keywords; [✂]; Reddit, 2026. Full automoderator documentation; Thorn, 2024. Introducing the CSAM Keyword Hub: A Free Collection of Words Related to Child Safety Risks.

[20] [✂]; Stream, 2026. Top Moderation Keyword Filters & Blocklists for 2026; Thorn, n.d. CSAM Keyword Hub.

Accuracy

A1.49 By definition, text matching approaches benefit from a high degree of transparency and explainability, as detections are generated on the basis of predefined indicators rather than inferred relationships.

A1.50 The degree of accuracy will depend on the quality and specificity of the indicators available to the system, as well as the extent to which those indicators are associated with the target harm. Where the quality and specificity are greater, the degree of accuracy is likely to be higher.

A1.51 Evidence indicates that distributors of image-based CSAM often use explicit CSAM-related terminology, coded language, abbreviations, identifiers and other textual signals when sharing, describing, searching for or facilitating access to such material.[22] These indicators may appear in captions, comments, filenames, metadata, usernames, URLs and associated communications.[23] Evidence further indicates that the identification of such terminology and coded language can be used to detect image-based CSAM and related offending activity.[24] We therefore consider that relevant textual indicators are sufficiently associated with the target harm to provide meaningful evidence of its likely presence.

A1.52 Lists of relevant indicators may be sourced from third-party intelligence initiatives, such as keyword datasets maintained by specialist organisations such as the Internet Watch Foundation Keywords List,[25] Thorn CSAM Keyword Hub[26] and Tech Coalition's Lantern initiative,[27] as well as being supplemented or developed internally based on the outcomes of content moderation processes. These keyword lists are developed globally and cover multiple languages, though there may be languages that are underrepresented.[28]

A1.53 The specialist datasets described above are maintained by organisations whose activities focus on combating online child sexual abuse and exploitation and which maintain processes for the ongoing review and updating of indicators. We consider that this specialist expertise, together with the governance and maintenance arrangements applied to these resources,[29] provide a reasonable basis for concluding that the datasets are of a high quality and suitable for use for the purpose of detecting unknown image-based CSAM.

---

[22] IWF, 2020. Significant breakthrough as IWF Analysts crack code used by predators to share child sexual abuse material online.

[23] Thorn, CSAM Keyword Hub; Westlake, B. and Guerra, E., 2023. Using file and folder naming and structure to improve automated detection of child sexual abuse images on the Dark Web, Forensic Science International: Digital Investigation, 47.

[24] In response to the June 2025 Consultation on Additional Safety Measures (June 2025 Consultation), the IWF said it provides a keyword list for its members to help them identify and take down illegal content on their services. IWF response to the June 2025 Consultation, p. 13. See also: Lykousas, N. and Patsakis, C., 2025. Just in plain sight: unveiling CSAM distribution campaigns on the clear web.; Kim, J. G. and Kim, J., 2026. Hybrid text mining for models for investigative keyword expansion on child sexual abuse in the dark web, PLoS One, 21 (5).; Wang Y., Arief, B., Franqueira, V. N. L., Coates, A. and Ciardha, C. Ó., 2023. Investigating the availability of child sexual abuse materials in dark web markets: evidence gathered and lessons learned, EICC '23: Proceedings of the 2023 European Interdisciplinary Cybersecurity Conference.

[25] IWF, n.d. Keywords.

[26] Thorn, 2024. Introducing the CSAM Keyword Hub: A Free Collection of Words Related to Child Safety Risks.

[27] Technology Coalition, n.d. Lantern: Advancing child safety through signal sharing.

[28] Thorn, Introducing the CSAM Keyword Hub: A Free Collection of Words Related to Child Safety Risks

[29] IWF, Keywords; Thorn, Introducing the CSAM Keyword Hub: A Free Collection of Words Related to Child Safety Risks

A1.54 The technical process of determining whether text matches a predefined keyword, phrase or pattern is well established.[30] However, different approaches may contribute to accuracy in different ways. Exact matching approaches can accurately identify predefined indicators where those indicators appear in the text being assessed. Pattern matching approaches can accurately identify text conforming to predefined patterns or structures, while similarity and approximate matching approaches can be configured to identify spelling variants, abbreviations, obfuscations and modified terminology that may not be captured through exact matching alone.[31] Such approaches are widely used in information retrieval, content moderation and safety systems, and their performance characteristics are measurable and configurable through the selection of matching rules and thresholds.

A1.55 Broader matching techniques may increase false positive risks because they identify text that is similar, rather than an exact match to, the target indicator. As similarity thresholds become broader, the likelihood of matching unrelated text may increase. However, this may also improve detection coverage by identifying relevant variations that would not be captured through exact matching alone.[32]

A1.56 Some residual false positive risk may remain even where datasets are of a high quality and matching approaches are appropriately configured. This is because particular keywords, phrases or patterns may be capable of both benign and harmful interpretation depending on the context in which they appear. Such risks may be mitigated through safeguards including the review of an appropriate proportion of matched content by human reviewers, in accordance with our draft guidance on human review in the Draft Proactive Technology Guidance, published alongside the statement.

A1.57 The proactive technology criteria recommend providers consider the quality, relevance and representativeness of datasets, evaluate performance using appropriate metrics, assess performance in real-world deployment contexts, and maintain mechanisms for ongoing review and improvement.[33] These criteria are intended to support identification and mitigation of limitations relating to indicator quality, coverage and detection performance.

A1.58 Taking these factors together, we consider that text matching approaches are capable of accurately detecting textual indicators associated with image-based CSAM. Given the demonstrated association between such indicators and the target harm, the availability of specialist and actively maintained indicator datasets, and the ability of text matching technologies to reliably identify those indicators, we consider that text matching is capable of supporting accurate detection outcomes when appropriately configured and maintained.

---

[30] Manning, C. D., Raghavan, P., & Schütze, H. 2008. Introduction to Information Retrieval (Vol. 39, pp. 234-265). Cambridge: Cambridge University Press.

[31] Manning, C. et al. Introduction to Information Retrieval.

[32] We recognise that in our December 2024 Statement, we decided not to recommend text matching for the detecting of CSAM because we considered that further evidence was needed about the quality, content (including specificity) and availability of existing keyword lists. We have decided to recommend this now having obtained more evidence and because we are now recommending this type of technology is only deployed where it meets the criteria, ensuring it is accurate, effective and lacking in bias.

[33] Proactive technology criteria of particular relevance include: (a) High Quality Data; (c) Evaluating Performance; (e) Contextual testing and evaluation; (f) Maintenance and Ongoing Monitoring; and (h) Incorporating Feedback.

Effectiveness

A1.59 Text matching approaches may be deployed across a range of service features, including content metadata, user-generated text, filenames, profile information, livestream chat and associated communications. This enables the technology to detect indicators of image-based CSAM across a range of content-sharing and content-discovery pathways.

A1.60 Where relevant indicators are present and accessible, text matching approaches can support effective detection through the identification of known terminology, patterns, phrases and other textual signals associated with image-based CSAM. The effectiveness of text matching nevertheless depends on the extent to which relevant textual indicators are present and available to the technology. Effectiveness will be reduced where image-based CSAM is not accompanied by textual indicators, or where relevant indicators are inaccessible to the provider.

A1.61 Maintaining effective detection depends upon the use of high-quality and appropriately maintained indicator datasets, together with processes for ongoing review and updating to reflect changes in terminology and patterns of abuse. Examples of such datasets include the Internet Watch Foundation's Keywords List and Thorn's CSAM Keyword Hub. We consider that these databases are compiled and maintained by experts in the identification of CSAM, with sufficient governance in place.[34] We understand that these databases are regularly reviewed and updated, with quality assurance processes in place.

A1.62 As discussed in the section entitled 'Accuracy', we consider these resources to provide a high-quality source of terminology and other indicators associated with CSAM, which may therefore be used to support the detection of image-based CSAM. Any dataset the provider uses for this purpose should meet the high-quality data criterion. For more detail on how to do this, see the section entitled 'Criterion A: High-Quality Data' in our Draft Proactive Technology Guidance, published alongside the statement.

A1.63 Where high-quality indicator datasets are used, and relevant indicators are present and accessible to the provider, text matching approaches are capable of reliably identifying those indicators within content.[35] The effectiveness of the technology therefore depends both on the quality and maintenance of the underlying indicator datasets and on the ability of the matching approach to accurately identify predefined keywords, phrases, patterns and related indicators when they occur.

A1.64 Pattern matching and similarity or approximate matching approaches may further improve effectiveness (depending on service context), by identifying variations, obfuscations and structured indicators that would not be detected through exact matching alone.

A1.65 Effectiveness may also be reduced where terminology evolves rapidly, indicators become outdated, or users deliberately seek to evade detection through obfuscation, abbreviations or alternative terminology. The quality, coverage and ongoing maintenance of the underlying indicator datasets are therefore important considerations in supporting effective detection.

A1.66 The proactive technology criteria recommend providers evaluate performance in real-world deployment scenarios, consider whether effectiveness may be improved through layering with complementary approaches, and maintain processes for ongoing monitoring,

[34] IWF, Keywords; and Thorn, Introducing the CSAM Keyword Hub: A Free Collection of Words Related to Child Safety Risks.

[35] Manning, C. et al. Introduction to Information Retrieval.

updating and maintenance.[36] These criteria are intended to support continued effectiveness where harmful behaviours, terminology or methods of evasion evolve over time.

A1.67 Taking these factors together, we consider that approaches within the text matching subclass are capable of contributing effectively to the purpose of detecting of unknown image-based CSAM across a range of deployment contexts.

Lack of bias

A1.68 As text matching relies upon predefined indicators, potential sources of bias are often more identifiable and explainable than in some other technology classes. As detections are generated from predefined indicators, providers are generally able to review, amend or remove indicators that give rise to unintended or disproportionate outcomes.

A1.69 However, bias risks may arise where datasets disproportionately reflect particular languages, communities, terminology or cultural contexts, or where relevant indicators are unavailable for specific user populations. These risks may affect all approaches within the subclass, including exact matching, pattern matching and similarity or approximate matching techniques.

A1.70 We consider that such risks are capable of mitigation through the development and maintenance of appropriately representative datasets, regular review processes, and consideration of the languages and user populations served by the relevant service provider.

A1.71 Further, as outlined above, specialist datasets are available to providers through organisations with expertise in combating online child sexual abuse and exploitation. Such specialist organisations apply governance, review and maintenance processes to support the ongoing identification, assessment and updating of relevant indicators. Where such datasets are used, these processes reduce the risk of systematic bias arising from outdated, inappropriate or insufficiently relevant indicators. While limitations may remain where relevant content is not represented within an indicator set, we consider these arrangements provide a reasonable basis for concluding that the risk of systematic bias arising from the indicators themselves is reduced.

A1.72 The proactive technology criteria recommend providers identify and address potential sources of bias, consider the suitability and representativeness of datasets and indicators, undertake appropriate evaluation activities, and maintain processes for ongoing review and improvement.[37] These criteria are intended to support identification and mitigation of systematic differences in outcomes that may arise during deployment.

Conclusion

A1.73 We have found that textual indicators associated with image-based CSAM are sufficiently associated with the target harm to support meaningful detection outcomes and that such indicators can be operationalised through appropriately configured text matching systems. We have also found that identified limitations may be mitigated through the use of appropriate indicator datasets, ongoing maintenance and deployment practices consistent

[36] Proactive technology criteria of particular relevance include: (e) Contextual testing and evaluation; (f) Maintenance and Ongoing Monitoring; (g) Human Review; and (h) Incorporating Feedback.

[37] Proactive Technology Criteria of particular relevance include: (a) High Quality Data; (b) Addressing Biases; (e) Contextual testing and evaluation; and (h) Incorporating Feedback.

with the proactive technology criteria. Our analysis shows that a number of high quality and well-maintained datasets containing these textual indicators are available.

A1.74 Taking these factors into account, we conclude that text matching technologies are capable in principle of being sufficiently accurate, effective and lacking in bias for the purpose of detecting unknown image-based CSAM where they meet the proactive technology criteria.

Application-specific machine learning for the purpose of detecting unknown image-based CSAM

A1.75 Application-specific machine learning are technologies based on models developed or adapted using training data to identify, classify, or assess features, attributes or characteristics of target content, and whose intended tasks, decision criteria and expected outputs are established through model development or adaptation.

A1.76 In the context of content identification, machine learning models are commonly used to classify content into predefined categories or generate information such as labels, scores or other structured outputs which can support moderation decisions. [38] [39]

A1.77 When performing image-analysis tasks, application-specific machine learning models analyse image data to identify the presence of characteristics associated with the target harm based on patterns learned during training. These patterns may include combinations of visual features, objects, scenes, textures, spatial relationships and other indicators associated with the target harm.

A1.78 In practice, application-specific machine learning implementations often rely upon convolutional neural networks (CNNs) or vision transformer (ViT) architectures, which are capable of identifying complex visual patterns within image content. [40]

A1.79 In the context of the detection of image-based CSAM, application-specific machine learning models operate through the direct analysis of image or video content (as relevant).

A1.80 Unlike text matching approaches, which identify textual indicators associated with the likely presence of the relevant harmful content, application-specific machine learning models evaluate the imagery itself and generate predictions about whether it exhibits characteristics associated with image-based CSAM.

A1.81 The core capabilities underpinning image and video analysis for this technology class are substantially similar from a technical perspective and are therefore assessed together below. Where additional considerations arise specifically in relation to video-based content, these are addressed separately within the relevant sections.

Accuracy

A1.82 Application-specific machine learning technologies are capable of identifying complex visual patterns that may be difficult to define through explicit rules alone. [41] A key strength

[38] [✂].

[39] Amazon, n.d. Using the image and video moderation APIs; Cambridge Consultants and Ofcom, 2019. Use of AI in online content moderation; Hive Moderation, n.d.; Microsoft, n.d. Content moderator API reference; and Winder.AI and Ofcom, 2023. Automated Content Classification (ACC) Systems.

[40] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Untherthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N., 2020 (revised 2021). An image is worth 16x16 words: transformers for image recognition at scale; Lecun, Y., Bottou, L., Bengio, Y. and Haffner, P., 1998. Gradient-based learning applied to document recognition, Proceedings of the IEEE, 86 (11); [✂].

[41] Codesota, 2026. ImageNet-1k, still the deciding test.

of the technology is that, by definition, it is able to generalise beyond examples encountered during training and identify previously unseen content which shares characteristics with known examples of the target content.

A1.83 Accuracy depends on a range of factors including the quality, quantity and representativeness of the training data,[42] the suitability of the model architecture,[43] calibration choices[44] and the extent to which content encountered during deployment reflects the content encountered during training and evaluation.[45]

A1.84 When applied to video content, application-specific machine learning systems may operate through analysis of individual frames, clips or other sampled visual sequences.[46] Accuracy may therefore depend upon factors such as sampling strategy,[47] aggregation approach[48] and the extent to which relevant indicators are observable within sampled content. For

---

[42] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child. R., Gray, S., Radford, A., Wu, J. and Amodei, D., 2020. Scaling laws for neural language models; Wiles, O., Gowal, S., Stimberg, F., Alvise-Rebuffi, S., Ktena, I., Dvijotham, K. and Cemgil, T., 2021. A fine-grained analysis on distribution shift; Klivans, A., Stavropoulos, K. and Vasilyan, A., 2024. Testable learning with distribution shift, Proceedings of Machine Learning Research, 196; Chen, M., Goel, K., Sohoni, N., Poms, F., Fatahalian, K. and Ré, C., 2021. Mandoline: Model evaluation under distribution shift, Proceedings of the 38th International Conference on Machine Learning, 139; Cai, T. T., Namkoong, H. and Yadlowsky, S., 2023. Diagnosing model performance under distribution shift; Mayo, D., Cummings, J., Lin, X., Gutfreund, D., Katz, B. and Barbu, A., 2023. How hard are computer vision datasets? Calibrating dataset difficulty to viewing time, Advances in Neural Information Processing Systems, 36 (NeurIIPS 2023); and Wang, C., Dong Q., Wang, X. and Sui, Z., 2022. Statistical dataset evaluation: reliability, difficulty and validity.

[43] Lakshminarayanan, B., Pritzel, A. and Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles; Richter, M. L., & Pal, C., 2022. Receptive field refinement for convolutional neural networks reliably improves predictive performance; Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N. and Polosukhin, I., 2017. Attention is all you need, Advances in Neural Information Processing Systems, 30.

[44] Guo, C., Pleiss, G., Sun, Y. and Weinberger, K. Q., 2017. On calibration of modern neural networks, ICML 2017.

[45] Barz, B. and Denzler, J., 2020. Do we train on test data? Purging CIFAR of near-duplicates, Journal of Imaging, 6 (6); Kaufman, S., Rosset, S., Perlich, C. and Stitelman, O., 2012. Leakage in data mining: Formulation, detection and avoidance, ACM Transactions on Knowledge Discovery from Data (TKDD), 6 (4).

[46] Pareek, P. and Thakkar, A., 2021. A survey on video-based Human Action Recognition: recent updates, datasets, challenges, and applications. Artificial Intelligence Review, 54 (3); Tang, Z., Zhao, Y., Wen, Y. and Liu, M., 2024. A survey on backbones for deep video action recognition; and Tran, D., Bourdev, L., Fergus, R., Torresani, L. and Paluri, M., 2015. Learning spatiotemporal features with 3D convolutional networks, Proceedings of the IEEE International Conference on Computer Vision (ICCV).

[47] Ding, H., Guo, C., Sun, J., Jiang, X., Shi, H. and Li, J., 2025. Motion-driven adaptive frame selection strategy for video action recognition, EURASIP Journal on Image and Video Processing, 2025 (12); Hu, K., Gao, F., Nie, X., Zhou, P., Tran, S., Neiman, T., Wang, L., Shah, M., Hamid, R., Yin, B. and Chilimbi, T., 2025. M-LLM based video frame selection for efficient video understanding, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2025); and Zhi, Y., Tong, Z., Wang, L. and Wu, G., 2021. MGSampler: an explainable sampling strategy for video action recognition, Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2021).

[48] Nam, H. and Park, Y, 2025. Temporal attention pooling for frequency dynamic convolution in sound event detection; Wang, J., Cherian, A., Porikli, F. and Gould, S., 2018. Video representation learning using discriminative pooling, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition 2018; Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X. and Van Gool, L., 2019. Temporal segment networks for action recognition in videos, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 41 (11); and Zou, J., Huang, Z., Zhang, S., Zhang, L. and Shen, W., 2026. VideoBrain: learning adaptive frame sampling for long video understanding.

example, some indicators of child sexual abuse may emerge over time rather than being visible within a single frame.

A1.85 Where application-specific machine learning technologies are applied to livestream content, accuracy may be affected by the need to assess content in near real time and by decisions regarding frame sampling, clip generation and processing frequency.[49] For example, relevant indicators of harm may appear only briefly within a livestream, meaning that detection performance may depend on the extent to which sampled content captures those indicators.

A1.86 Where representative training data is available, application-specific machine learning technologies are capable of achieving high levels of classification performance across image-classification tasks.[50] However, performance may be reduced where harmful content is rare, highly contextual, rapidly evolving or underrepresented within training datasets.

A1.87 Machine learning approaches are widely used for image classification and content identification tasks, including the detection of image-based CSAM and related harms.[51]

A1.88 Commercial providers, trust and safety vendors and non-profit organisations offer machine-learning-based solutions designed to support the identification of previously unknown image-based CSAM.[52]

A1.89 We recognise that access to representative training and evaluation datasets presents a particular challenge for unknown image-based CSAM. However, evidence demonstrates that the National Center for Missing and Exploited Children (NCMEC) and other specialist organisations[53] have facilitated access to representative CSAM datasets or related training resources under tightly controlled governance arrangements.[54]

A1.90 The existence of these arrangements provides evidence that representative datasets are available to support the development, training and evaluation of machine learning models for this use case. We consider that the specialist expertise of the organisations responsible for these resources, together with the governance arrangements applied to them, provide a reasonable basis for concluding that the datasets are of a high quality and suitable for use in training and evaluating machine learning systems designed to detect unknown image-based CSAM.

A1.91 In some deployments, application-specific machine learning technologies may be trained directly on datasets containing examples of CSAM. In other deployments, models may instead be trained to identify visual characteristics associated with CSAM, such as indicators relating to age, nudity, sexual activity or other contextual features. We consider that both approaches fall within the application-specific machine learning technology class

---

[49] Ding et al. Motion-driven adaptive frame selection strategy for video action recognition; Hu et al. M-LLM based video frame selection for efficient video understanding; and Zhi et al. MGSampler: an explainable sampling strategy for video action recognition.

[50] Codesota, ImageNet-1k; [✂]; Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S. and Hu, H., 2021. Video swin transformer.; [✂]; and [✂].

[51] [✂]; [✂]; [✂]; [✂]; [✂]; [✂]; Thorn, 2025, CSAM has distinct characteristics that call for purpose-built solutions; and Thorn, n.d. Solutions for content-hosting platforms.

[52] Alice, n.d. Activefence – User generated content; [✂]; Google, n.d. Tools for partners; [✂]; [✂] Thorn, Solutions for content-hosting platforms; and WebPurify, n.d. aim.imgcheck.

[53] Roke, n.d. Vigil AI; and Thorn, 2026. Safer Predict overview.

[54] Thorn, 2025. Announcing Safer Predict: AI-driven CSAM & CSE Detection.

and may be capable of supporting the detection of image-based CSAM, provided they are appropriately trained, evaluated and deployed to meet the proactive technology criteria.

A1.92 The proactive technology criteria ask providers to consider the suitability and representativeness of training and evaluation data, undertake testing and validation activities, and maintain ongoing monitoring processes.[55] These criteria are intended to support identification and mitigation of such limitations throughout the deployment lifecycle.

A1.93 As set out in this section, where representative training is available, application-specific machine learning technologies are capable of achieving high levels of classification performance across image-classification tasks. As we have explained, in the case of image-based CSAM, representative training data is available. Taking this and the other factors discussed in this section into account, we consider that application-specific machine learning is capable of being accurate at detecting image-based CSAM when trained, calibrated and maintained in a manner that meets the proactive technology criteria.

Effectiveness

A1.94 A key strength of machine learning models is scalability. Once deployed, models are capable of assessing very large volumes of content and can operate continuously across substantial quantities of image content in a manner that would not be achievable through human review alone.

A1.95 Effectiveness depends not only on the performance of the model itself but also on supporting processes such as data collection, annotation, retraining, monitoring and performance evaluation. It may also be affected by concept drift,[56] changing patterns of harmful behaviour and deliberate attempts to evade detection.[57]

A1.96 Machine learning approaches are therefore most effective where representative training datasets are available and where ongoing monitoring and model maintenance allow systems to adapt to changes in content over time. As noted above, we know that organisations such as NCMEC have facilitated the availability of representative training datasets related to image-based CSAM. [58]

A1.97 Machine learning approaches are already deployed in production moderation systems to support the identification of image-based harms at scale.[59]

A1.98 The availability of commercial products, third-party moderation services and internally developed provider solutions indicates that machine learning approaches are capable of

---

[55] Proactive Technology Criteria of particular relevance include: (a) High Quality Data; (c) Evaluating Performance; (e) Contextual testing and evaluation; (f) Maintenance and Ongoing Monitoring; and (h) Incorporating Feedback.

[56] Goodfellow, I. J., Mirza, M., Xiao, D., Courville, A. and Bengio, Y., 2013. An empirical investigation of catastrophic forgetting in gradient-based neural networks; Serra-Perello, M. and Ortiz, A., 2025. Incremental learning methodologies for addressing catastrophic forgetting: analysis and experimentational evaluation, Journal of Artificial Intelligence Research, 83; and van de Ven, G. M., Soures, N. and Kudithipudi, D., 2025. Continual learning and catastrophic forgetting.

[57] Goodfellow, I. J., Shlens, J. and Szegedy, C., 2015. Explaining and Harnessing Adversarial Examples; and Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. and Fergus, R., 2014. Intriguing Properties of Neural Networks.

[58] Roke, Vigil AI; and Thorn, Safer Predict overview.

[59] [✂]; [✂]; [✂]; [✂]; [✂]; [✂]; Thorn, CSAM has distinct characteristics that call for purpose-built solutions; and Thorn, Solutions for content-hosting platforms.

operating within real-world moderation workflows.[60] Providers commonly deploy such systems as part of wider moderation processes involving monitoring, review and model maintenance.[61]

A1.99 Video classification presents additional technical challenges, including decisions relating to frame sampling, clip generation, temporal analysis and latency requirements. Effectiveness may therefore be influenced by choices regarding sampling frequency, aggregation techniques, latency tolerances and livestream processing requirements.[62] However, these largely represent implementation considerations rather than inherent barriers to the use of machine learning technologies for image and video classification tasks.

A1.100 When deployed against livestream content, effectiveness may also be influenced by latency requirements and available processing resources.[63] However, application-specific machine learning approaches are capable of being deployed within livestream moderation workflows and may be used to analyse livestream content for visual characteristics associated with image-based CSAM during or shortly after transmission, depending on the deployment model adopted.[64]

A1.101 The existence of a mature ecosystem of machine learning tools and providers provides evidence that application-specific machine learning is operationally deployable and capable of contributing to the detection of unknown image-based CSAM under real-world conditions.[65]

A1.102 Further, the proactive technology criteria recommend providers monitor performance on an ongoing basis and consider whether technologies remain effective in the deployment context in which they are used. This includes consideration of retraining, updating and maintenance activities designed to address changing patterns of harmful behaviour over time.[66]

A1.103 Taking these factors together, we consider that technologies within the application-specific machine learning class are capable of being effective for the purpose of detecting unknown image-based CSAM at scale.

---

[60] Alice, Activefence – User generated content; [✂]; Google, Tools for partners; [✂]; [✂]; [✂]; [✂]; Thorn, CSAM has distinct characteristics that call for purpose-built solutions; Thorn, Solutions for content-hosting platforms; and WebPurify, aim.imgcheck

[61] [✂]; [✂]; [✂]; and [✂]

[62] Ding et al. Motion-driven adaptive frame selection strategy for video action recognition; Hu et al. M-LLM based video frame selection for efficient video understanding; Nam and Park. Temporal attention pooling for frequency dynamic convolution in sound event detection; Wang et al. Video representation learning using discriminative pooling; Wang et al. Temporal segment networks for action recognition in videos; Zhi et al. MGSampler: an explainable sampling strategy for video action recognition; and Zou,et al. VideoBrain: learning adaptive frame sampling for long video understanding.

[63] Visive.ai, 2026. Why latency matters in human-AI video; Yew et al. Dynamic Content Moderation in Livestreams: Combining Supervised Classification with MLLM-Boosted Similarity Matching

[64] Zhang L. and Vu, T., 2023. Moderate your Amazon IVS live stream using Amazon Rekognition, AWS Blogs.

[65] CaseScan (n.d.) CaseScan CSAM Alternative; Google (n.d.) Tools for Partners; Resolver (n.d.) CSAM Detection; Roke (n.d.) Vigil AI: Protecting Children Against Online Harms Using AI; Safer (n.d.) Solutions; The Hive AI (n.d.) CSAM Detection AP.

[66] Proactive Technology Criteria of particular relevance include: (f) Maintenance and Ongoing Monitoring; and (h) Incorporating Feedback.

Lack of bias

A1.104 Bias risks are well recognised within machine learning development and evaluation processes and are therefore subject to significant ongoing research and governance activity.[67]

A1.105 Bias risks may arise where training or evaluation data systematically underrepresent particular groups, contexts or content characteristics.[68] Model performance may also vary where the content encountered in operation differs materially from the data used during training and evaluation.[69]

A1.106 In the context of image-based CSAM detection, bias risks may arise through demographic imbalances within training datasets, differences in image quality, or the model learning spurious visual characteristics that are not genuinely associated with the target harm. These risks are capable of mitigation through the use of representative datasets, careful dataset governance, robust evaluation processes, ongoing monitoring and regular model retraining where appropriate.

A1.107 Similar considerations apply in video-based deployments, where any demographic or content imbalances present in the underlying training and evaluation datasets may also influence performance.

A1.108 Like text-matching, we recognise that access to datasets presents a particular challenge for unknown image-based CSAM. Available datasets are typically derived from content which has already been identified and reported and may not fully reflect the wider distribution of harmful content encountered in practice.

A1.109 Accordingly, model performance depends not only on access to training data, but also on the quality, representativeness and governance of those datasets, together with the approaches used to address potential imbalances and sources of bias during training and evaluation.

A1.110 However, specialist organisations such as NCMEC have facilitated access to relevant datasets and training resources under tightly controlled governance arrangements. In other words, appropriately governed datasets are available to support development, training and evaluation of machine learning approaches for this use case.

A1.111 The proactive technology criteria also recommend providers consider sources of systematic error and bias, undertake appropriate testing and evaluation, and maintain governance

[67] Information technology – Artificial intelligence – Treatment of unwanted bias in classification and regression machine learning tasks; Information technology – artificial intelligence (AI) – Bias in AI systems and AI aided decision making; Software and systems engineering – Software testing, Part 11: Guidelines on the testing of AI-based systems; Information technology – Artificial intelligence – Management system; Information technology – Artificial intelligence – Assessment of machine learning classification performance.

[68] Buolamwini, J. and Gebru, T., 2018. Gender shades: intersectional accuracy disparities in commercial gender classification, Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR, 81.

[69] Cai, Namkoong and Yadlowsky. Diagnosing model performance under distribution shift; Chen, et al. Mandoline: Model evaluation under distribution shift; and Klivans, Stavropoulos, and Vasilyan. Testable learning with distribution shift.

processes designed to identify and address material impacts on different groups, contexts and content types.[70]

A1.112 Taking these factors together, our analysis concludes that identified bias risks are capable of mitigation through testing, evaluation and deployment practices consistent with those set out within the proactive technology criteria.

Conclusion

A1.113 Application-specific machine learning technologies are capable of directly analysing image and video content and identifying visual features associated with image-based CSAM. Where representative training data is available, such technologies are capable of achieving high levels of classification performance across image-classification tasks. While access to such datasets can present challenges in the case of CSAM, specialist organisations such as NCMEC have facilitated access to representative CSAM datasets and related training resources. We also consider that the limitations associated with these approaches can be mitigated through appropriate dataset governance, calibration, testing, monitoring and maintenance practices.

A1.114 Taking these factors together, we conclude that application-specific machine learning technologies are capable in principle of accurately and effectively supporting detection of image-based CSAM, in a manner that lacks undue bias, where they meet the proactive technology criteria.

Foundation Models

A1.115 Within the Foundation Models technology class our analysis identified two subclasses: Large Language Models (LLMs) and Multimodal Foundation Models (MFMs).

A1.116 Our analysis found that conceptually, there are grounds to consider that both LLMs and MFMs may be capable of supporting the detection of image-based CSAM.

A1.117 In the case of MFMs, their ability to analyse visual and contextual information simultaneously suggests that they may be capable of identifying characteristics associated with the target harm across multiple sources of information, including previously unseen image and video content.

A1.118 In the case of LLMs, they are capable of identifying semantic meaning, contextual relationships and implied references that may not be captured through predefined indicators alone.[71] As a result, they may be capable of identifying and interpreting coded language, understanding contextual relationships between terms, and identifying indicators of harmful content that may not be represented within predefined keyword sets.

A1.119 We also identified evidence that foundation model technologies are increasingly being deployed within content moderation, trust and safety and content analysis workflows. This provides evidence that the technology class is operationally deployable and capable of supporting moderation activities.

---

[70] Proactive Technology Criteria of particular relevance include: (b) Addressing biases; (c) Evaluating Performance; (e) Contextual testing and evaluation; (f) Maintenance and Ongoing Monitoring; and (h) Incorporating Feedback.

[71] Bommasani et al, 2022. On the Opportunities and Risks of Foundation Models; Vaswani et al, 2023. Attention Is All You Need; Brown et al, 2020. Language Models are Few-Shot Learners.

A1.120 However, whilst we identified evidence regarding the use of foundation model technologies within moderation and trust and safety contexts more generally,[72] we identified comparatively limited evidence regarding their operational deployment and evaluation specifically for the purpose of detecting unknown image-based CSAM.

A1.121 Taking this into account, we do not currently consider there to be a sufficient evidential basis to recommend foundation model technologies within our measure on the deployment of proactive technology to detect image-based CSAM content (ICU PT1). This does not prevent providers from using foundation model technologies for this use case should they choose to do so. It remains open to service providers to use this technology if they wish to, as a matter of commercial choice.

A1.122 We recognise that both foundation model technologies and their application to harmful content detection in trust and safety environments are evolving rapidly. We will therefore continue to monitor developments in this area and may consider foundation model technologies further as part of future work.

Conclusion

A1.123 We consider that text matching and application-specific machine learning technologies are each capable in principle of accurately and effectively supporting the detection of image-based CSAM, in a manner that lacks undue bias. While each technology class exhibits distinct strengths, limitations and deployment considerations, the available evidence indicates that those limitations are capable of mitigation where they meet the proactive technology criteria.

[72] Open AI, 2024. Upgrading the Moderation API with our new multimodal moderation model; Google Cloud (Gemini Enterprise Agent Platform), 2026. Gemini for safety filtering and content moderation; Arxiv, 2023. Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations; Meta, n.d. Llama Protections - Llama; Reddit for Developer, 2026. AI AutoModerator.

A2. Proactive technology: draft guidance changes

Reasoning for changes to the Draft Proactive Technology Guidance

Ensuring the guidance provides sufficient specificity and foreseeability

A2.1 Stakeholders expressed concern that several aspects of the draft guidance were insufficiently specific and afforded providers broad discretion in the implementation of the proactive technology criteria. In particular, stakeholders questioned whether certain criteria were clear and foreseeable, including the appropriate balance between precision and recall, the appropriate proportion of outputs for human review and the acceptable rate of false positives.[73] Some stakeholders were concerned that the approach taken in the draft guidance could lead to inconsistent application between providers, making it difficult for users to predict how content would be treated in practice.[74]

A2.2 Stakeholders also noted that false positives are an inevitable outcome of proactive technologies and sought greater clarity regarding how providers should identify, assess and respond to them. In particular, stakeholders questioned whether providers would be expected to minimise false positives or also take steps to identify and correct false positives where they arise.[75]

---

[73] A member of the Philippine Survivor Network response to the June 2025 Additional Safety Measures Consultation, pp. 11-12; BILETA response to the June 2025 Additional Safety Measures Consultation, pp. 8-9; Gender and Tech Research Lab's response to the June 2025 Additional Safety Measures Consultation, p. 4; Internet Watch Foundation response to the June 2025 Additional Safety Measures Consultation, p. 18, 23; Microsoft response to the June 2025 Additional Safety Measures Consultation, p.3; [✂]; National Society for the Prevention of Cruelty to Children (NSPCC) response to the June 2025 Additional Safety Measures Consultation, pp. 2-4, 26; NPCC response to the June 2025 Additional Safety Measures Consultation, p. 4; Open Rights Group response to the June 2025 Additional Safety Measures Consultation, pp. 16-17, 19-20, 25; Professor Korff, D. response to the June 2025 Additional Safety Measures Consultation, p. 4, 6. 'Fully-automated free speech-killing robots are incompatible with the European and UK fundamental rights'.; Respondents to the Open Rights Group Campaign response to the June 2025 Additional Safety Measures Consultation, pp. 2, 5-6; Smith, G. response to the June 2025 Additional Safety Measures Consultation, pp. 1-9. 'Ofcom's proactive technology measures: principles-based or vague?'.; South West Grid for Learning response to the June 2025 Additional Safety Measures Consultation, p. 7; [✂]; [✂]; The Commissioner Designate for Victims of Crime Northern Ireland response to the June 2025 Additional Safety Measures Consultation, p. 5; UK Finance response to the June 2025 Additional Safety Measures Consultation, p. 9; [✂].

[74] A member of the Philippine Survivor Network response to the June 2025 Consultation, pp. 11-12; [✂]; Professor Korff, D. response to the June 2025 Consultation, p. 4, 6; Respondents to the Open Rights Group Campaign response to the June 2025 Consultation, pp. 2, 5-6; Smith, G. response to the June 2025Consultation, pp. 1-9; [✂].

[75] Are, C. Dr response to the June 2025 Additional Safety Measures Consultation, p. 15; BILETA response to the June 2025 Consultation, p. 8; Domestic Abuse Commissioner response to the June 2025 Additional Safety Measures Consultation, p. 5; Evangelical Alliance's response to the June 2025 Additional Safety Measures Consultation, p. 3; Glitch response to the June 2025 Additional Safety Measures Consultation, p. 3; Information Commission's Office (ICO) response to the June 2025 Additional Safety Measures Consultation, p.10; Marie

A2.3 To address these concerns, we have made a number of changes to the draft guidance to provide greater specificity regarding the implementation and operation of the measure on the deployment of proactive technology to detect image-based CSAM content. These changes are intended to improve the predictability and foreseeability of the measure, support more consistent implementation by providers, and provide greater clarity regarding our expectations of providers while retaining flexibility to account for differences in service design and risk profiles. The changes made in response to these concerns are identified in Table A2.1.

Reducing the risk of incorrect moderation decisions and prior-restraint

A2.4 Stakeholders raised concerns about the risk of incorrect moderation decisions resulting from the use of proactive technology, including the potential for lawful content to be wrongly detected, restricted or removed. Several stakeholders highlighted the risks associated with automated decision making, particularly where content moderation decisions may be taken before content is published or without sufficient human review. This included concerns about compatibility with automated decision-making requirements under UK GDPR.[76]

A2.5 To address these concerns, we have strengthened the guidance relating to a number of the proactive technology criteria. These changes are intended to reduce the likelihood and impact of incorrect moderation decisions, including where proactive technologies may limit the availability of content before users are able to engage with it, while preserving providers' ability to act swiftly to address harm. The changes made in response to these concerns are identified in Table A2.1.

Reducing the risk of unfairness and bias

A2.6 Stakeholders also expressed concerns that automated systems may struggle to account for context, linguistic nuances and minority languages, increasing the risk of erroneous decisions and disproportionate impacts on particular groups of users.[77]

A2.7 Stakeholders emphasised the importance of safeguards, including human review, as well as the data protection principles of fairness, accuracy and transparency.[78]

A2.8 To address these concerns, we have strengthened the draft guidance relating to a number of the proactive technology criteria. These changes are intended to reduce the risk of bias and strengthen protections for user rights. The changes made in response to these concerns are identified in Table A2.1.

Collins Foundation response to the June 2025 Additional Safety Measures Consultation, p. 7; Name Withheld 4 response to the June 2025 Additional Safety Measures Consultation, pp. 3-4; Online Responsibility Network response to the June 2025 Additional Safety Measures Consultation, p. 3; Open Rights Group response to the June 2025 Consultation, pp. 14-15; Professor Korff, D. response to the June 2025 Consultation, pp. 11-12; Respondents to the Open Rights Group Campaign response to the June 2025 Consultation, p. 2; Smith, G. response to the June 2025 Consultation, p. 2; [✂].

[76] Cook, J. response to the June 2025 Additional Safety Measures Consultation, p. 2; [✂]; [✂]; Open Rights Group response to the June 2025 Consultation, pp. 25-26; Professor Korff, D. response to the June 2025 Consultation, pp. 5, 8-9; Smith, G. response to the June 2025 Consultation, p. 5; [✂].

[77] [✂]; Glitch response to the June 2025 Consultation, pp. 11-12; NSPCC response to the June 2025 Consultation, pp. 52-54; Open Rights Group response to the June 2025 Consultation, pp. 45-46; [✂].

[78] ICO response to the June 2025 Consultation, p. 8; Professor Korff, D. response to the June 2025 Consultation, pp. 9-10.

Reducing the risk of chilling effects

A2.9 Stakeholders raised concerns that the use of proactive technology could have a chilling effect on users' willingness to engage in lawful expression online because of fear their content will be incorrectly detected or removed.[79]

A2.10 To address these concerns, we have strengthened the guidance in relation to a number of the proactive technology criteria. These changes are intended to improve the consistency and predictability of moderation outcomes. The changes made in response to these concerns are identified in Table A2.1.

Other changes

A2.11 In addition to the changes made for the reasons set out above, we have made a number of further amendments to the draft Proactive Technology Guidance. These amendments include the following changes, which align with the amendments to the measure itself:

a) Recommending the deployment, rather than the assessment, of specified technologies,[80] together with associated amendments explaining that providers should develop proactive technology in-house where suitable third-party solutions are not available;

b) Recommending the use of specified technologies for the purpose of detecting unknown image-based CSAM content, rather than a broader set of harms; and

c) Clarifying that providers may choose to use a combination of specified technologies.

A2.12 These changes were made to ensure that the draft guidance more accurately reflects the scope of the measure and the practical ways in which providers may implement it. The reasons for these changes are explained in chapter 9: Deployment of proactive technology to detect image-based CSAM of the statement but are summarised in Table A2.1 for completeness.

Changes to the Draft Proactive Technology Guidance

A2.13 Table A2.1 summarises the substantive changes, the reasons for those changes and the stakeholder concerns they are intended to address. In some cases, a single amendment addresses more than one concern and therefore contributes to multiple objectives identified above. The table does not set out the amendments to the wording of the proactive technology criteria as the rationale for those changes is already explained in chapter 9: Deployment of proactive technology to detect image-based CSAM of the statement.

---

[79] BILETA response to the June 2025 Consultation, p. 8; Cook, J. response to the June 2025 Consultation, p. 2; European Digital Rights Organisation response to the June 2025 Additional Safety Measures Consultation, p. 2; [✂]; Name Withheld 4 response to the June 2025 Consultation, p. 3; Name Withheld 22 response to the June 2025 Additional Safety Measures Consultation, p. 3; Professor Korff, D. response to the June 2025 Consultation, p. 6; Respondents to the Open Rights Group Campaign response to the June 2025 Consultation, p. 2; Respondents to the Together Declaration Campaign response to the June 2025 Additional Safety Measures Consultation, p. 4-5; [✂]; The Commissioner Designate for Victims of Crime Northern Ireland response to the June 2025 Consultation, p. 5; [✂]; [✂].

[80] Where we assessed a class and/or subclass of technology is capable in principle of being accurate, effective and lacking in bias, where it meets our proactive technology criteria, we refer to these technologies as 'specified technologies'.

Table A2.1: Table of changes to the draft Proactive Technology Guidance

Section of the draft guidance	Change made	Reason for Change
Various	Amended the draft guidance to reflect the change to the measure from assessing proactive technology to selecting and using specified technologies to meet the proactive technology criteria.	Please see chapter 9: Deployment of proactive technology to detect image-based CSAM.
Various	Amended references from service's risk assessment to provider's risk assessment.	Please see the section entitled 'High-quality data' in chapter 9: Deployment of proactive technology to detect image-based CSAM content.
Various	Clarification that providers may choose to use a combination of the specified technologies.	Please see chapter 9: Deployment of proactive technology to detect image-based CSAM.
Various	Added references to the lifecycle of the relevant technology.	To clarify when activities set out in the proactive technology criteria apply at different stages in the lifecycle of the relevant technology, recognising that some activities are relevant throughout the entire lifecycle while others apply only at particular stages.
'ICU PT 1 - Deployment of proactive technology to detect image-based CSAM content'	Clarification that, where their own existing proactive technology cannot be configured in a way that meets the proactive technology criteria, providers should source from a third party, develop in-house or a combination of these approaches.	Please see chapter 9: Deployment of proactive technology to detect image-based CSAM.

Section of the draft guidance	Change made	Reason for Change
'ICU PT 1 - Deployment of proactive technology to detect image-based CSAM content'	Amended the guidance to include the records providers should make.	To set out the kinds of records which need to be kept in order for a provider to know that it has implemented and maintained proactive technology consistently with the proactive technology criteria. Lack of appropriate records impacts user safety because it prevents a service provider from being able to understand and evaluate its own decisions, monitor its progress and learn from past experience.
'Interaction with data protection law'	Clarification that providers should continue to refer to the relevant guidance issued by the Information Commission's Office (ICO) to ensure they are complying with their data protection obligations.	To clarify the main data protection principles relevant to the use of proactive technology and that providers should stay up to date with ICO's guidance on these topics.
'The proactive technology criteria'	Added further explanation of the concepts of accuracy, effectiveness and lack of bias.	To explain more clearly how Ofcom's guidance to service providers on accuracy, effectiveness and lack of bias relates to the statutory use of those terms.
'The criteria'	Clarification that the protective technology criteria form part of an ongoing and interconnected process.	To ensure the draft guidance provides sufficient specificity and foreseeability and to reduce the risk of incorrect moderation decisions.
'The process of applying the proactive technology criteria'	Strengthened guidance on obtaining legal advice relating to relevant criminal offences and defences and referred providers to the Memorandum of Understanding between the Crown Prosecution Service and the National Police Chiefs' Council.	To support providers in implementing the measure lawfully and with confidence.
'Initial assessment'	Clarification that providers that decide to develop proactive technology in-house do not need to conduct the initial assessment stage of the process. Added factors providers should initially consider where developing proactive technology in-house.	To ensure the draft guidance provides sufficient specificity and foreseeability.

Section of the draft guidance	Change made	Reason for Change
'Understanding the proactive technology criteria'	Clarification that providers should consider how the harm manifests on their service, as identified through their risk assessment, when considering how to configure and deploy proactive technology in a way that meets the criteria.	To clarify that providers should take account of how image-based CSAM manifests on their service, as identified through their risk assessment, when configuring and deploying proactive technology, helping to ensure the technology is effective in practice at detecting and addressing the harm.
A. High-quality data	Strengthened guidance for providers to assess whether datasets used are of sufficient high-quality, to identify available mitigations to address gaps or limitations, and regarding the level of expertise required to assess and quality assure datasets.	To ensure the draft guidance provides sufficient specificity and foreseeability, and to reduce the risks of incorrect moderation decisions, unfairness and bias.
B. Addressing biases	Strengthened guidance on identifying, monitoring and mitigating bias that the relevant technology has toward protected characteristics under the Equality Act 2010 and referred providers to the guidance from the Information Commission's Office on the safeguards required for processing special category data in automated decision-making.	To ensure the draft guidance provides sufficient specificity and foreseeability, and to reduce the risks of incorrect moderation decisions, prior-restraint, chilling effects, unfairness and bias.
C. Evaluating performance	Clarified the objective when determining the appropriate balance between precision and recall and that decisions should take account of potential impacts on user rights, including privacy or legitimate political expression.	To ensure the draft guidance provides sufficient specificity and foreseeability, and to reduce the risks of incorrect moderation decisions, prior-restraint, chilling effects, unfairness and bias.

Section of the draft guidance	Change made	Reason for Change
E. Contextual testing and evaluation	Strengthened guidance for providers when determining the reliability of detection outputs, their level of assurance in the performance of the relevant technology and circumstances in which their level of assurance should be treated as being materially reduced.	To ensure the draft guidance provides sufficient specificity and foreseeability, and to reduce the risks of incorrect moderation decisions, prior-restraint, chilling effects, unfairness and bias.
F. Maintenance and ongoing monitoring	Added quality assurance as part of the maintenance and ongoing monitoring of the relevant technology, and clarified the role of quality assurance and how providers should conduct it.	To ensure the draft guidance provides sufficient specificity and foreseeability, and to reduce the risks of incorrect moderation decisions, prior-restraint, chilling effects, unfairness and bias.
G. Human review	Clarified the distinction between human review for moderation purposes and quality assurance purposes and that quality assurance includes the review of undetected content, clarified that human review supports compliance with data protection legislation and strengthened the guidance on determining the appropriate proportion of human review.	To ensure the draft guidance provides sufficient specificity and foreseeability, and to reduce the risks of incorrect moderation decisions, prior-restraint, chilling effects, unfairness and bias. Further reasoning explaining the changes to the human review criterion are set out in Annex 4.3: Detailed reasoning on human review.
H. Incorporating feedback	Strengthened guidance on obtaining and assessing feedback, identifying false positives and false negatives, responding to identified errors, updating the relevant technology, re-evaluating performance following changes and records which should be kept.	To ensure the draft guidance provides sufficient specificity and foreseeability, and to reduce the risks of incorrect moderation decisions, prior-restraint, chilling effects, unfairness and bias.

A3. Human review for proactive technology

Introduction

A3.1 In the June 2025 Consultation, stakeholders provided feedback on the appropriate role and use of human review in the proposed measures on proactive technology assessed by the provider. In response to that feedback, we have clarified the purposes and appropriate proportion of human review under measure ICU PT1 (Deployment of proactive technology to detect image-based child sexual abuse material (CSAM) content). This annex explains those clarifications and our reasons for making them.

A3.2 Our reasoning in relation to human review under the measure on the deployment of proactive technology to detect image-based CSAM content is substantively the same as our reasoning on human review under the hash matching measures. We considered this stakeholder feedback when finalising our May 2026 Statement recommending the use of hash matching to detect intimate image abuse content (May 2026 Statement) and set out our position on the appropriate proportion of human review of detected intimate image abuse content.[81]

A3.3 Having adopted that position in relation to the intimate image abuse hash matching measure, we have considered its applicability to our other hash matching measures. Accordingly, in this annex we also set out our position on the human review of an appropriate proportion of detected content in relation to both the hash matching for terrorism content measure (ICU C13) and the hash matching for CSAM measure (ICU C9).

A3.4 For the purposes of the measures, there are two types of human review:

a) For moderation purposes (namely, to make decisions about individual items of content); and
b) For quality assurance purposes (namely, to test, monitor and improve the accuracy of moderation tooling and processes).

A3.5 In the section entitled 'Feedback on human review', we summarise the stakeholder feedback received during the June 2025 Consultation on human review and explain how it has informed our approach.

Ofcom's consultation proposals

Proactive technology criterion: 'Human review'

A3.6 In the June 2025 Consultation, we proposed human review as part of the proactive technology criteria in the draft Illegal Content Codes of Practice. This criterion stated that "policies and processes are in place for human review and action is taken in accordance with that policy, including the evaluation of outputs during development (where applicable), and the human review of an appropriate proportion of the outputs of the proactive technology during deployment. Outputs should be explainable to the extent necessary to support meaningful human judgment and accountability."

[81] Ofcom's Codes and Guidance following that Statement were issued on 9 September 2026.

Determining the appropriate proportion for human review

A3.7 In the June 2025 Consultation, we further proposed a number of factors to be taken into account when determining the 'appropriate proportion' of outputs of the proactive technology for human review under the measure on proactive technology assessed by the provider, namely:

a) the principle that the resource dedicated to review of detected content should be proportionate to the degree of accuracy achieved by the technology and any associated systems and processes;
b) the principle that content with a higher likelihood of being a false positive should be prioritised for review; and
c) in the case of CSAM or grooming, the importance of minimising the reporting of false positives to the National Crime Agency (NCA) or a foreign agency.

A3.8 We also proposed a number of factors to be taken into account when determining the 'appropriate proportion' of detected terrorism content to be subject to human review under the measure on using hash matching to detect terrorism content, namely:

a) The degree of accuracy achieved by automated systems and/or processes in use for the purposes of content moderation, if any (taking account of data from the service's complaints procedures);
b) The principle that content with a higher likelihood of being a false positive should be prioritised for review; and
c) The importance of understanding the purpose and context of detected content when determining whether it is terrorism content.

A3.9 The first two of the factors with respect to the measure on hash matching to detect terrorism content, and all three of the factors with respect to the measure on proactive technology assessed by the provider, drew from our existing Code measure on hash matching for CSAM (ICU C9). This sets out that the factors to be taken into account when determining the appropriate proportion of human review in the measure on using hash matching for CSAM are:

a) The principle that the resource dedicated to review of detected content should be proportionate to the degree of accuracy achieved by the perceptual hash matching technology in use and any associated systems and/or processes (as indicated by the periodic reviews of the performance of the technology mentioned in measure ICU C9.13, and also taking account of the outcomes of reviews of content carried out by human moderators and data from the service's complaints procedure enabling United Kingdom users to complain if content they have generated, uploaded or shared is taken down on the basis that it is illegal content);
b) the principle that content with a higher likelihood of being a false positive should be prioritised for review; and
c) the importance of minimising the reporting of false positives to the NCA or a foreign agency (within the meaning of Chapter 2 of Part 4 of the Act).

Feedback on human review

A3.10 We received several responses from stakeholders regarding human review of detected content within our proposed measures on proactive technology assessed by the provider. In particular, these responses concerned the circumstances in which human review should be used, the risks associated with fully automated moderation, the impact of incorrect

detections on users and the need for clear and foreseeable safeguards. Some of the responses submitted in respect of our proposals on providers assessing proactive technology raised issues which are equally relevant to the use of human review within the context of hash matching.

A3.11 We have therefore considered these responses when finalising the measure on the deployment of proactive technology to detect image-based CSAM content and the accompanying draft proactive technology guidance, as well as the measures on perceptual hash matching for terrorism content and hash matching for CSAM, and the accompanying guidance on human review for hash matching.

A3.12 Several stakeholders argued that our proposed factors governing human review did not provide sufficient certainty regarding when detected content should be reviewed by human moderators and left too much discretion to service providers.[82] They suggested that providers could reach different conclusions regarding the proportion of detected content that should be subject to human review. They considered this could result in inconsistent implementation across services and reduce the predictability and foreseeability of the regime for users.[83]

A3.13 Some stakeholders also expressed concerns regarding the use of automated systems to review content detected by proactive technologies.[84] They argued that content could be suppressed through fully automated processes without human oversight which amounts to prior constraint on freedom of expression,[85] while other stakeholders raised concerns that increased reliance on automated detection could result in lawful content being incorrectly removed[86] or a breach of the automated decision-making requirements and fairness principle under data protection legislation.[87] Several stakeholders emphasised the importance of human review in improving the accuracy and reliability of hash matching of terrorism content, particularly by tackling false positives, and preventing infringements on freedom of expression.[88] Two stakeholders emphasised the need for human reviewers to be trained in order to act as the main safeguard against bias and removal of lawful speech.[89]

---

[82] Open Rights Group response to the June 2025 Additional Safety Measures Consultation, p. 25; Professor Korff, D. response to the June 2025 Additional Safety Measures Consultation, p. 4. 'Fully-automated free speech-killing robots are incompatible with the European and UK fundamental rights'; Smith, G. response to the June 2025 Additional Safety Measures Consultation, pp. 4-6. 'Ofcom's proactive technology measures: principles-based or vague?'; The Open Rights Group Campaign response to the June 2025 Consultation, p.5.

[83] A member of the Philippine Survivor Network response to the June 2025 Additional Safety Measures Consultation, p. 11; Open Rights Group response to the June 2025 Consultation, p. 25; Professor Korff, D. response to the June 2025 Consultation, p.4; Smith, G. response to the June 2025 Consultation, pp. 5-6; The Open Rights Group Campaign response to the June 2025 Consultation, p.5.

[84] [✂]; Professor Korff, D. response to the June 2025 Consultation, pp. 8-9.

[85] Professor Korff, D. response to the June 2025 Consultation, pp. 8-9.

[86] [✂]; Open Rights Group response to the June 2025 Consultation, p. 25; Professor Korff, D. response to the June 2025 Consultation, pp. 8-9.

[87] Information Commission's Office (ICO) response to the June 2025 Additional Safety Measures Consultation, p. 3, 10 & 14; Professor Korff, D. response to the June 2025 Consultation, p. 9.

[88] British and Irish Law Education Technology Association (BILETA) response to the June 2025 Consultation on Additional Safety Measures, pp. 16-17; Open Rights Group response to June 2025 Consultation on Additional Safety Measures pp. 28-29. [✂]; Tech against terrorism response to the June 2025 Consultation p. 8.

[89] BILETA response to the June 2025 Consultation p. 17; Computer and Communications Industry Association (CCIA) response to the June 2025 Additional Safety Measures Consultation, p. 10.

A3.14 A number of stakeholders welcomed the use of human review as a safeguard within content moderation systems but sought greater clarity regarding how it should operate in practice. Some stakeholders emphasised the importance of clear expectations regarding moderation standards,[90] while others stated that the number of human moderators should be proportionate to the risk of harm on the service.[91] Stakeholders also highlighted the importance of appropriately trained and resourced human moderation for complex or high-risk decisions.[92]

A3.15 Stakeholders also commented on how providers should prioritise content for review. One service provider suggested that human review processes should prioritise content which is more likely to be violative over content which is more likely to result in a false positive,[93] while others noted that providers should consider the consequences of incorrect detections, including the impact of moderation action on users, when determining the appropriate level of human review.[94]

A3.16 Several stakeholders emphasised the need for human review to operate as an effective safeguard against incorrect moderation decisions and highlighted the importance of considering the consequences of moderation outcomes.[95] In particular, the Information Commission's Office (ICO) suggested that providers should consider the impact and consequences of incorrect detections when determining the appropriate level of human review.[96] Some stakeholders highlighted the integration of human review as a form of best practice.[97]

Our decision

Proactive technology criterion: 'Human review'

A3.17 In the measure 'Deployment of proactive technology to detect image-based CSAM content' (ICU PT1), within the 'Human review' criterion we have clarified that human review serves two purposes, namely as part of moderation decisions and for quality assurance. We have also clarified when human review should be undertaken. For moderation purposes, providers should ensure the human review of an appropriate proportion of outputs of the

[90] Marie Collins Foundation response to the June 2025 Additional Safety Measures Consultation, p. 7; NSPCC response to the June 2025 Additional Safety Measures Consultation, pp. 2-4; South West Grid for Learning response to the June 2025 Additional Safety Measures Consultation, p. 7; The Commissioner Designate for Victims of Crime Northern Ireland response to the June 2025 Additional Safety Measures Consultation, p. 4; The Domestic Abuse Commissioner response to the June 2025 Additional Safety Measures Consultation, p. 5.
[91] Marie Collins Foundation response to the June 2025 Consultation, p. 7; NSPCC response to the June 2025 Consultation, pp. 2-4; South West Grid for Learning response to the June 2025 Consultation, p. 7; The Commissioner Designate for Victims of Crime Northern Ireland response to the June 2025 Consultation, p. 4; The Domestic Abuse Commissioner response to the June 2025 Consultation, p. 5.
[92] Marie Collins Foundation response to the June 2025 Consultation, p. 7; The Domestic Abuse Commissioner response to the June 2025 Consultation, p. 5
[93] Snap Inc response to the June 2025 Additional Safety Measures Consultation, p. 6.
[94] ICO response to the June 2025 Consultation, p. 9.
[95] Marie Collins Foundation response to the June 2025 Consultation, p. 7; NSPCC response to the June 2025 Consultation, pp. 2-4; South West Grid for Learning response to the June 2025 Consultation, p. 7; The Domestic Abuse Commissioner response to the June 2025 Consultation, p. 5.
[96] ICO response to the June 2025 Consultation, p. 10.
[97] Snap, Inc. response to the June 2025 Consultation, p. 10; [✂].

proactive technology during deployment. For quality assurance purposes, providers should ensure human review throughout the lifecycle of the relevant technology.[98]

A3.18 We have made these clarifications in response to stakeholder feedback seeking greater clarity regarding the role of human review, including when it should take place and the safeguards it is intended to provide. By distinguishing between human review undertaken for moderation purposes and human review undertaken for quality assurance purposes, we have clarified the different functions that human review serves within the deployment and monitoring of relevant technology. We have also specified when each type of review should occur. Requiring human review for moderation purposes during deployment helps ensure that providers maintain appropriate human oversight of automated outputs when taking moderation action. Requiring human review for quality assurance purposes throughout the lifecycle of the relevant technology helps providers assess and maintain the performance, accuracy and reliability of the relevant technology over time. Taken together, these changes provide greater clarity and consistency regarding the operation of human review.

A3.19 We have also clarified that human review undertaken for quality assurance purposes requires reviewing an appropriate proportion of undetected content. For example, content identified as being suspected target content through user reporting, complaints, community moderation, stratified sampling or other means. This clarification reflects the role of quality assurance in assessing the ability and effectiveness of the relevant technology for the purpose of detecting unknown image-based CSAM content. Reviewing an appropriate proportion of undetected content helps providers to identify instances where image-based CSAM content has not been detected by the proactive technology (false negatives) and assess whether adjustments to the technology are required.

A3.20 We consider this amendment responds to stakeholder feedback seeking greater certainty about the role and purpose of human review within the measure. Clarifying that quality assurance human review includes the review of undetected content provides greater clarity on how human review can support ongoing assessment of the proactive technology's performance. By including undetected content within the human review criterion, this helps providers identify both content incorrectly detected (false positives) and also content that has not been detected when it should have been (false negatives).

A3.21 We have not made an equivalent amendment to the hash matching measures. Within the long-established structure of the Code measures concerned, the factors relevant to determining the appropriate proportion of human review relate only to matches that have been detected by the hash matching technology. However, in configuring the hash matching technology so that its performance strikes an appropriate balance between precision and recall, we do expect providers to undertake appropriate quality assurance activities to ensure effective use of the hash matching technology. This includes, where appropriate, human review of content that has not been detected by the hash matching technology. Such review can assist providers in identifying false negatives and assessing whether changes to the configuration of the technology are required to ensure an appropriate balance between precision and recall.

[98] A form of proactive technology that constitutes one or a combination of the specified technologies for the purposes of ICU PT 1 and has been implemented by a provider to detect target content. Where we assessed a class and/or subclass of technology is capable in principle of being accurate, effective and lacking in bias, where it meets our proactive technology criteria, we refer to these technologies as ‘specified technologies’.

Determining the appropriate proportion of human review

Ofcom's May 2026 Statement on hash matching for intimate image abuse

A3.22 In light of the feedback from stakeholders as outlined in the section entitled 'Feedback on human review', we reviewed our approach to the appropriate proportion of human review when we issued our May 2026 Statement. Measures ICU C14 and ICS C8 provide that, in determining the appropriate proportion of human review for content detected using hash matching for intimate image abuse, providers should have regard to the following factors:

a) the level of assurance the provider has in the detection outcomes achieved by the hash matching technology and any associated systems and processes (as indicated by the ongoing monitoring, evaluation and quality assurance (including human quality assurance) of the performance of the technology); and
b) to the extent that detected content is subject to review for the purpose of Recommendation ICU C1:
    i) the potential severity of the harm to those depicted in or encountering intimate image abuse content; and
    ii) the overall impact of an incorrect decision that the relevant content is illegal content on a United Kingdom user who generated, uploaded or shared the content.

A3.23 We explained that in our view, it is not possible to be prescriptive about the proportion of detected content that should be subjected to human review. Rather, the proportion of detected content should depend on the factors we have set out above. For example, all other things being equal, where the provider has configured the perceptual hash matching technology to prioritise recall there is a greater risk of false positives and therefore a greater proportion of human review is needed to mitigate this. If the prevalence of intimate image abuse content on a service is very low, then even very accurate technology could result in a large amount of content being detected which is not intimate image abuse content. We expect that service providers will adjust the configuration of their technology or have human moderators review a higher proportion of matches, if they observe that lawful content is being detected. Regardless of how providers configure their technology, we consider that some human review is always appropriate for quality assurance purposes. We therefore consider that some human review is always appropriate, but the extent of it should depend on the factors identified in our measure.

A3.24 As to the level of assurance a service provider has in the detection outcomes, we explained that this is closely related to the degree of accuracy achieved by automated systems and/or processes and the likelihood of false positive detections. Where the technology and associated systems and/or processes are highly accurate, the risk of false positives is generally lower. Conversely, where there is greater uncertainty as to the performance of the technology, the likelihood of false positives increases, and a higher proportion of human review will likely be appropriate. Indicators such as proximity to a matching threshold, the use of verified or unverified hash databases, and observed error rates should be understood collectively as signals of assurance rather than as standalone factors, informing the extent to which providers can rely on automated outputs without further verification. Service providers are expected to conduct ongoing monitoring and quality assurance (including human review) of the hash matching technology, to assess both the accuracy of the technology and relevant systems and processes, and the prevalence of incorrect detections over time. A service provider should take the outcomes of this monitoring and quality assurance into account when determining its level of assurance in the detection outcomes.

A3.25 The appropriate level of human review depends not only on the performance of the hash matching technology but also on the effectiveness of other technical and human review layers. While further review may be warranted where content is detected, it is not necessarily appropriate to prioritise human review solely based on a higher likelihood of false positives. Providers should instead consider how human judgment can meaningfully add value beyond automated systems and prioritise review where it is most likely to affect content moderation outcomes.

A3.26 Increased quality assurance, including human review where appropriate, is particularly important where assurance is lower or where observed outcomes are inconsistent with expectations.

A3.27 As to severity of harm, we noted that human review of content usually takes more time than automated processes do. We consider that the appropriate proportion of human review for moderation purposes is affected by the potential severity of the harm. If, for example, a provider is noticing that intimate image abuse content is frequently going viral, or a misogynistic culture is developing on its service, it might be appropriate to reduce the proportion of content which is subject to human review for moderation purposes temporarily in order to address the problem.

A3.28 We consider service providers should consider the impact of an incorrect decision that the content is illegal content on the posting user or interested person when determining the proportion of detected content for human review. This is because the harm caused by an incorrect decision will vary.

Appropriate proportion of human review across the proactive technology and hash matching measures

A3.29 We have also revised the factors affecting the "appropriate proportion of human review" for our measures on the deployment of proactive technology to detect image-based CSAM content and hash matching for both terrorism content and CSAM. We consider that the reasoning we laid out in relation to intimate image abuse in our May 2026 Statement is broadly relevant for each of these measures, and we explain this in more detail in this annex.[99] We have also taken the view that it is desirable for the hash matching measures for different harms to be aligned in their wording, where relevant.

A3.30 Therefore, for our measure on the deployment of proactive technology to detect image-based CSAM content, we have:

a) Combined and revised the first two existing factors to "the level of assurance the provider has in the detection outcomes achieved by the technology and any associated systems and processes".
b) Added a new factor: the severity of harm to those depicted in or encountering content.
c) Widened the factor on the importance of minimising the reporting of false positives to the NCA or a foreign agency to include other impacts and clarified that this factor is relevant for moderation purposes rather than quality assurance purposes.

A3.31 Therefore, for our measure on hash matching for CSAM, we have:

a) Combined and revised the first two existing factors to "the level of assurance the provider has in the detection outcomes achieved by the technology and any associated systems and processes".

[99] Ofcom (2026), Statement - Detecting Intimate Image Abuse, paragraphs 3.166-3.176.

b) Added a new factor for moderation purposes: severity of harm to those depicted in or encountering content.
c) Widened the factor on the importance of minimising the reporting of false positives to the NCA or a foreign agency to include other impacts and clarified that this factor is relevant for moderation purposes rather than quality assurance purposes.

A3.32 This change applies to all services in scope of the CSAM hash matching measure.

A3.33 For our measure on hash matching for terrorism content, we have:

a) Combined and revised the first three factors on which we consulted to "the level of assurance the provider has in the detection outcomes achieved by the technology and any associated systems and processes."

b) Added new factors for moderation purposes: Severity of harm to those depicted in or encountering content and overall impact of an incorrect decision on the posting user.

A3.34 A comparison of the factors for each of the hash matching measures is set out below.

A3.35 For our measure on the deployment of proactive technology to detect image-based CSAM content, and for all the harms for which we have recommended the use of hash matching, providers should, when determining the appropriate proportion of human review for moderation and quality assurance purposes, have regard to the level of assurance they have in the detection outcomes achieved by the technology and any associated systems and processes. Their level of assurance will be indicated by the ongoing monitoring, evaluation and quality assurance (including human quality assurance) of the performance of the technology.

A3.36 In addition, to the extent that detected content is subject to review for the purpose of Recommendation ICU C1 (i.e. for moderation purposes), a provider should have regard to:

<table>
<tr><th>Intimate image abuse</th><th>Terrorism</th><th>CSAM</th></tr>
<tr><td>the potential severity of the harm to those depicted in or encountering intimate image abuse content; and</td><td>the potential severity of the harm to those depicted in or encountering terrorism content; and</td><td>the potential severity of the harm to those depicted in or encountering CSAM; and</td></tr>
<tr><td colspan="2">the overall impact of an incorrect decision that the relevant content is illegal content on a United Kingdom user who generated, uploaded or shared the content.</td><td>the overall impact of an incorrect decision that the relevant content is illegal content on a United Kingdom user who generated, uploaded or shared the content, including the importance of minimising the reporting of false positives to the National Crime Agency (NCA) or a foreign agency (within the meaning of Chapter 2 of Part 4 of the Act).</td></tr>
</table>

A3.37 For the measure on the deployment of proactive technology to detect image-based CSAM content, we have included guidance which sets out how to apply all the factors in more detail within the Draft Proactive Technology Guidance, published alongside the statement.

A3.38 We have also decided to issue similar draft guidance on human review in respect of our measures on hash matching to detect CSAM and terrorism content, and we have consolidated this with our existing guidance on intimate image abuse.[100] The full draft guidance, which we will issue when we issue the Codes, has been published alongside this statement.

A3.39 Both these sets of draft guidance clarify the role of human review for both moderation and quality assurance purposes, explain the relevance of levels of assurance in detection outcomes, and provide further detail regarding circumstances in which increased human review is likely to be appropriate.

A3.40 We explain our reasoning in relation to the factors affecting the "appropriate proportion of human review" for our measures on the deployment of proactive technology to detect image-based CSAM content and hash matching for both terrorism and CSAM, in more detail below.

Moderation purposes

What we mean by "the level of assurance the provider has in the detection outcomes achieved by the technology and any associated systems and processes"

A3.41 The level of assurance a provider has in the detection outcomes is closely related to the degree of accuracy achieved by automated systems and/or processes and the likelihood of false positive detections. Where the technology and associated systems and/or processes are highly accurate, the risk of false positives is generally lower. Conversely, where there is greater uncertainty as to the performance of the technology, the likelihood of false positives increases, and a higher proportion of human review will likely be appropriate.

A3.42 We recognise that if the prevalence of the illegal content concerned on a service is very low, even highly accurate technology may generate a considerable proportion of false positives. This is known as a 'false positive paradox'. Where a false positive paradox arises, a provider will need to consider the increased likelihood that a positive detection could be a false positive.

A3.43 However, hash matching for CSAM can be operated to achieve very high levels of assurance in the detection outcomes, and further assurance can be added by layering with other kinds of technology. For example, a provider might use more than one perceptual hash matching algorithm, which can provide additional certainty that identified matches are accurate and reduce the likelihood that a particular result is attributable to a false positive in a single algorithm); or use machine-learning classifiers to identify items of detected content that are more or less likely to be false positives.

A3.44 By contrast, we expect that service providers are unlikely to be able to achieve high levels of assurance in detection outcomes for most terrorism content and will need human review of most if not all detected items. This is because most terrorism offences are contextual. They require consideration of the state of mind of the person posting the content and/or its

[100] The wording of our intimate image abuse guidance has not changed but we have moved one paragraph for clarity.

effects on users likely to encounter it. At present we are not aware of automated technologies which are capable of carrying out this sort of analysis accurately, though we recognise that technology is continually developing.

A3.45 An exception to this general rule may arise the second time an item of content is matched if the service provider has retained a record from the first time it was reviewed, that the content amounts to one of the few terrorism offences which do not need to be considered in context. We consider this to be the case where the content amounts to:

a) the offence of collection of information likely to be of use to a terrorist; or
b) instruction or training in the making or use of radioactive material, weapons designed or adapted for the discharge of any radioactive material, explosives, chemical, biological or nuclear weapons. This includes blueprints for printing 3D firearms.

A3.46 With respect to the proactive technology measure for the deployment of proactive technology to detect image-based CSAM content and the hash matching measures, where the provider has a lower level of assurance in the detection outcomes produced by the technology, we expect a higher proportion of content to be subject to human review.

A3.47 As for intimate image abuse hash matching, we have decided not to retain a standalone factor requiring providers to prioritise content which is more likely to be a false positive. The appropriate level of human review depends not only on the performance of the technology but also on the effectiveness of other technical and human review layers. While further review may be warranted where content is detected, it is not necessarily appropriate to prioritise human review solely based on a higher likelihood of false positives. Providers should instead consider how human judgment can meaningfully add value beyond automated systems and prioritise review where it is most likely to affect content moderation outcomes.

A3.48 A service provider which has chosen to configure its technology with broad parameters (for example, reflective of its risk appetite or in order to ensure that its content moderators are always occupied), and which allows its users to encounter content pending human review, should not necessarily prioritise review of content which is unlikely to be illegal, over content which is likely to be illegal.

New factor – severity of harm to those depicted in or encountering content

A3.49 Human review of content usually takes more time than automated processes do. We consider that the appropriate proportion of human review for moderation purposes is affected by the potential severity of the harm.

A3.50 This may change from time to time. If, for example, a provider is noticing a significant uptick of terrorism or CSAM content on its service generally (whether detected through proactive technology, hash matching or in other ways), it might be appropriate to reduce the proportion of content which is subject to human review for moderation purposes temporarily in order to address the problem.

A3.51 We consider that CSAM is severely harmful both to those depicted in the content and those who may encounter it. As a result, it may be appropriate for providers to take down content detected as CSAM on a precautionary basis, without prior human review. With respect to hash matching, this could apply even where the provider has deployed relatively broad perceptual hash matching parameters and the content produces a match that falls towards the lower end of the matching threshold.

New factor – overall impact of an incorrect decision on the posting user

A3.52 The consequences of a false positive may vary significantly depending on the action taken against the user and the extent to which decisions can be reversed. For example, some users may not be inconvenienced much by content take-down, but those monetising their content may experience a loss of income. Decisions which lead to user banning have a greater impact than decisions which only lead to content take-down.

A3.53 We therefore consider that the likely impact on the posting user is relevant when determining the appropriate proportion of human review. This helps ensure that human review resources are directed towards cases where the consequences of an incorrect decision are most significant.[101]

A3.54 When content is taken down from the service without human review on a precautionary basis, it may still be appropriate to undertake a subsequent human review of the content before imposing a more serious sanction on a user, depending on the provider's level of assurance in the correctness of the detection outcome.

Quality assurance purposes

A3.55 For the measure on the 'Deployment of proactive technology to detect image-based CSAM content (ICU PT1), service providers are expected to conduct ongoing monitoring and quality assurance as part of the proactive technology criteria 'Maintenance and ongoing monitoring' and 'Incorporating feedback'. In conducting these processes, human review of detected content for quality assurance purposes may involve reassessing a sample of detected content and content moderation decisions relating to such content (including both automated and human decisions). Quality assurance of proactive technology also involves the human review of undetected content, as set out in the section entitled 'Proactive technology criterion: Human review' of this chapter.

A3.56 For both hash matching measures, service providers are expected to conduct ongoing monitoring and quality assurance (including human review) of their hash matching technology and any associated moderation systems and processes. This includes an assessment both of the accuracy of the technology and relevant systems and processes, and the prevalence of incorrect detections over time.

A3.57 Regardless of how providers configure their hash matching technology, we consider that some human review is always appropriate for quality assurance purposes but the extent of it should depend on the factors identified in our measure.

A3.58 For our measures on proactive technology and hash matching technologies, providers should take the outcomes of this monitoring and quality assurance into account when determining its level of assurance in the detection outcomes.

A3.59 Increased quality assurance, including human review where appropriate, is particularly important where assurance is lower or where observed outcomes are inconsistent with expectations for our measures on proactive technology and hash matching technologies.

[101] ICO response to the June 2025 Consultation, p. 9; Marie Collins Foundation response to the June 2025 Consultation, p. 7; The Domestic Abuse Commissioner response to the June 2025 Consultation, p. 5.

Impact assessments

A3.60 Our impact assessment of the measure on the 'Deployment of proactive technology for detecting image-based CSAM content' (ICU PT1) is included in chapter 9: Deployment of proactive technology to detect image-based CSAM.

A3.61 Our impact assessment of the measure on hash matching for terrorism content is included in chapter 10: Hash matching.

A3.62 The changes to the measure on hash matching for CSAM are a change to the existing measure. However, we consider this amendment to be a clarification rather than a substantive change to the measure. Accordingly, we do not expect it to result in providers requiring additional human review resources or otherwise giving rise to significant additional costs.

Conclusion

A3.63 Overall, we consider that it is not appropriate to be prescriptive about the proportion of detected content that should be subjected to human review for our measures on proactive technology and hash matching. For our measure on the deployment of proactive technology to detect image-based CSAM content, we also consider that it is not appropriate to be prescriptive about the proportion of undetected content that should be subjected to human review for quality assurance purposes. However, we consider our amended factors, interpreted in accordance with the accompanying guidance, will steer service providers appropriately to weigh up risks of harms against impacts on freedom of expression and the interests of users who are acting lawfully.

A4. Hash matching for terrorism content: detailed reasoning on effectiveness

A4.1 This annex sets out our detailed reasoning for believing that hash-matching deployed in accordance with our recommendation will be accurate, effective and lacking in bias in detecting suspected terrorism content. It sets out our further reasoning on:

a) Effectiveness of perceptual hash matching over cryptographic hash matching,
b) Availability of a hash database that meets our criteria for "an appropriate set of hashes",
c) Appropriate configuration of perceptual hash matching technology, and
d) Revision of terminology: replacing "visual images (whether or not combined with written material)" with "visual materials".

Effectiveness of perceptual hash matching over cryptographic hash matching

A4.2 Some respondents noted the benefits of cryptographic hash matching.[102] One service provider explained that the effectiveness of hash matching is highest for known material that has already been flagged and hashed. It is less suited to identifying entirely new or modified content.[103]

A4.3 We consider that the benefits of this measure will be maximised if service providers use perceptual hash matching rather than cryptographic hash matching. This is because perceptual hashing can identify content that has been slightly changed, not just exact copies. Perceptual hash matching is capable of detecting visually similar content, making it better suited to limiting the repeated sharing of this content at scale.[104] If we were to recommend relying solely on cryptographic hashing, users posting illegal terrorism content could readily circumvent the protections provided by this measure by making minor modifications to the images they upload, thereby avoiding detection.

A4.4 The ability of perceptual hash matching technology to identify both exact copies and slightly altered versions of content enables it to detect a wider range of harmful material than cryptographic hash matching. This broader detection capability makes it more effective in limiting the re-upload and continued spread of terrorism content.

[102] Google LLC response to the June 2025 Consultation, p.21; Tech Against Terrorism response to the June 2025 consultation; [✂].
[103] OnlyFans response to the June 2025 Consultation, email correspondence.
[104] Global Internet Forum to Counter Terrorism (GIFCT), 2023. Advances in Hashing for Counterterrorism. 29 March.; McDonald, B., 2022. Extremists are seeping back into the mainstream: Algorithmic detection and evasion tactics on social media platforms, Global Network on Extremism and Technology, 31 October; Macdonald, S., Mattheis, A. and Wells, D., 2024. Using Artificial Intelligence and Machine Learning to Identify Terrorist Content Online, Tech Against Terrorism Europe, pp.16-17.

Availability of a hash database that meets our criteria for "an appropriate set of hashes"

A4.5 The effectiveness of this measure and the benefits it delivers are directly linked to the provider's choice of hash database. This is because the coverage, quality and accuracy of the underlying hashes directly determine how much terrorism content can be detected and how accurately it is identified.

A4.6 As explained in the section titled 'Our decision' in chapter 10: Hash matching – Measure on perceptual hash matching for terrorism content, we have decided that if a service provider does not have access to an appropriate set of hashes developed in-house, it should use a third-party dataset that meets our criteria. We explain our new expectations and the reasons for this change in the following sections, along with the criteria for an appropriate set of hashes and how those criteria have been updated.

A4.7 We have also concluded that we needed to set clearer expectations about what counts as a suitable hash database, specifically what we mean by "an appropriate set of hashes". We explain this in the section titled 'Updated criteria for "an appropriate set of hashes"' below.

Service providers should use a third party database where they do not have an internal appropriate set of hashes (database)

A4.8 For the measure to be effective, the providers of all services in scope need ongoing access to an appropriate set of hashes. We therefore consider that where a service does not have such a set, it should source hashes from a third-party database that meets our criteria. This ensures that all in-scope services have access to hashes that are accurate and up to date, while recognising that no existing dataset fully aligns with the offences set out in the Online Safety Act 2023 (the Act).

A4.9 Where a service provider relies on a third-party dataset, it should also incorporate hashes generated through its own content moderation activities to ensure material identified on the service is captured and not re-uploaded. Such hashes should be incorporated as soon as reasonably practicable.

A4.10 We have taken this approach because developing an appropriate terrorism hash database internally can take significant time, during which risks to users may remain unaddressed. Using a suitable third-party dataset is therefore a more effective option, particularly for smaller services. It also allows service providers to benefit from existing expertise and maintain reliable datasets.

A4.11 This decision has been shaped by stakeholder feedback. Several stakeholders raised concerns about the quality of hash datasets, noting that the effectiveness of hash matching depends heavily on dataset quality.[105] Some stakeholders referred to existing third-party hash databases as part of their responses.[106] Some respondents flagged concerns with third party datasets.[107] We explore what constitutes an appropriate set of hashes, be it third party and/or internally generated, in the following sections.

---

[105] Mega response to the June 2025 Consultation, pp.19-20; Middle Tech Coalition response to the June 2025 Consultation, p.4; Pinterest response to the June 2025 Consultation, p.5; techUK response to the June 2025 Consultation, pp.15-16.

[106] GitHub response to the June 2025 Consultation, p.6; OSA Network response to the June 2025 Consultation, p.20; Tech Against Terrorism response to the June 2025 Consultation, pp.7-8.

[107] CCIA response to the June 2025 Consultation, p.10; Free Speech Union response to the June 2025 Consultation, p.5; Google LLC response to the June 2025 Consultation, p.7; [✂]; [✂].

Updated criteria for "an appropriate set of hashes"

A4.12 To support providers in sourcing an appropriate hash database, we set out the updated criteria in the following paragraphs and explain the reasons for each change in turn.

A4.13 We have decided to simplify the drafting of the measure. We have merged our expectations on "a person with expertise" with our criteria for "an appropriate set of hashes".

A4.14 As outlined in the section titled 'How this measure works' in chapter 10: Hash matching – Measure on perceptual hash matching for terrorism content, we have set out updated criteria to support providers in sourcing an appropriate hash database. We review these each in turn in the following paragraphs and explain our reasoning.

Criterion A – Include hashes of terrorism content sourced by a person with expertise in identification of terrorism content

A4.15 In the June 2025 Additional Safety Measures Consultation (June 2025 Consultation), we proposed that an appropriate set of hashes should include terrorism content identified either through a provider's own moderation processes or by one or more persons with expertise in identifying such content.[108] However, we consider that the criteria for an "appropriate set of hashes" and the expectations placed on a person with relevant expertise are effectively the same. We have therefore simplified the drafting by removing the separate reference to a "person with expertise" and incorporating those expectations into the definition of an appropriate set of hashes. This is not a substantive change to the measure, but a clarification intended to reduce complexity and avoid confusion.

Criterion B – Include an appropriate range of content amounting to terrorism

A4.16 We have decided to keep the criterion that hashes should reflect the range of content that amounts to an offence under Schedule 5 of the Act. To be effective in protecting users, a database must cover the breadth of relevant offences, rather than a narrow subset – for example, a database that was limited to information likely to be of use to a terrorist would not necessarily capture the ranges of offences to do with proscribed organisations. We also mean that an appropriate database should not be limited to any single type of terrorism, such as far-right terrorism or Islamist terrorism, but should instead include a range of terrorist ideologies.

A4.17 We have decided to remove references to risk assessments from this criterion. As we are now providing for the use of suitable third-party datasets, we no longer consider it appropriate to refer to a service's risk assessments or other internal information in the criteria. We consider that terrorism content is not generally service-specific in a way that always requires a separate dataset for each service. A dataset that reflects the range of relevant terrorism offences can therefore help effectively identify terrorism content, regardless of whether it was compiled with reference to a particular service's risk profile. Any matched content can then be assessed by the service provider in context through its moderation processes. Third-party providers are also unlikely to have access to a service's risk assessments or the outcomes of its content moderation processes.

[108] A "person with expertise in the identification of terrorism content" may be the service provider's content moderation staff, an expert third party database provider, or other expert in the identification of terrorism content. This person should have arrangements in place to identify suspected terrorism content and to ensure that terrorism content is correctly identified before hashes of that material are added to the database.

Criterion C – Contain hashes of a significant number of original items of content

A4.18 We have introduced an additional criterion that a dataset should include hashes relating to a sufficiently large volume of original terrorism content. This is a matter of logic – to be effective, a database must cover a significant amount of content; a database derived from a small number of content items would not provide meaningful mitigation.

A4.19 In terms of quantity, the dataset should be at least comparable to established third-party databases used to identify terrorism content.

Criterion D: Does not plainly discriminate on the basis of protected characteristics (within the meaning of Part 2 of the Equality Act 2010)

A4.20 Respondents raised concerns about the risk of bias in terrorism hash databases, particularly the potential for disproportionate impacts on users communicating in certain languages or belonging to specific cultural, national or religious groups.[109]

A4.21 We recognise this risk, but we do not consider it necessary to make further changes to the measure. This is because we already proposed that the arrangements in place should not plainly discriminate against groups with protected characteristics, as set out at [ICU C13.8(c)/13.10(b)].

A4.22 Where datasets are compiled by persons with expertise in identifying terrorism content (whether internally or through a third-party provider), any bias is more likely to result in the omission of some terrorism content rather than the incorrect inclusion of non-terrorism material. In the UK, a greater number of Islamist terrorist organisations have historically been proscribed than other types of organisations. As a result, a higher proportion of Islamist-related content in a dataset is more likely to reflect the scope of the law and the length of time for which these organisations have been proscribed, than a bias in its compilation.

A4.23 In the unlikely event that content is incorrectly identified by a suitably expert third-party provider, we recognise a theoretical risk of disproportionate impact on certain groups, including those with protected characteristics. However, we consider that the expertise and processes of such providers are sufficient to mitigate this risk.

A4.24 This underscores the importance of safeguards to minimise impacts on lawful expression and protect users' rights, which we address further in the section titled 'Rights impact' in chapter 10: Hash matching – Measure on perceptual hash matching for terrorism content.

Criterion E: Be proactively updated with reasonable regularity

A4.25 In our June 2025 Consultation, we explained that providers should regularly obtain updated versions of any hash databases, as databases that are not maintained risk becoming out of date. We have reflected this in our updated criteria of an appropriate set of hashes, which now includes the expectation that databases are proactively updated with reasonable regularity. This recognises that effective hash matching depends on capturing new and modified content as the harm evolves and is informed by feedback from the Integrity Institute highlighting the practical challenges of keeping datasets up to date.[110]

---

[109] BILETA response to the June 2025 Consultation p.16, p.6; Open Rights Group response to the June 2025 Consultation, p.29; Digital Resilience in Education – Welsh Government response to the June Consultation, p 10.

[110] Integrity Institute response to the June 2025 Consultation, p.6.

Criterion F: Be subject to arrangements for identifying and removing hashes that are not capable of amounting to terrorism content

A4.26 We have clarified that the service provider should take reasonable steps to secure the removal of a hash from the hash database where the provider has determined that the hash is not capable of amounting to terrorism content (for example, it is a picture of a family pet). We have also explained that where a service provider uses a third-party database, it will satisfy this criterion by notifying the hash database provider of the incorrect hash.

Criterion G: Be secured from unauthorised access or interference

A4.27 For a database to be accurate, it needs to be secure from unauthorised access or interference. We have therefore updated the criteria for an appropriate set of hashes to include an expectation that databases are protected from unauthorised access or interference, whether by internal staff, service providers, or external actors. This applies to both service providers' internal databases and third-party datasets. Providers are expected to secure their own systems and take reasonable steps to ensure that any third-party datasets they use are similarly safeguarded. This criterion is intended to preserve the integrity and reliability of hash datasets, reduce the risk of inaccurate or malicious hashes being introduced, and protect freedom of expression.

Hashing of new content detected on the service via content moderation

A4.28 Service providers should add hashes to their database when they become aware of new terrorism content on their service, either through user reporting, user complaints or other means. This aspect of the measure falls outside the criteria for an appropriate set of hashes because third-party dataset providers cannot see or access a provider's internal content moderation decisions. As a result, only the provider can update its dataset using hashes of content that it has identified through its own moderation processes.

A4.29 We have made the decision to maintain this recommendation because, for a hash database to be accurate and effective at identifying terrorism content, it should include content that a service provider has already identified through its own moderation processes. Excluding such content would create gaps in coverage, reducing the system's overall effectiveness. While we consider third-party datasets that meet our criteria to be sufficient, particularly for smaller services that may lack the resources to build their own databases from nothing, we do not consider it sufficient for service providers to rely solely on third-party datasets and ignore the outcomes of their own moderation decisions. Failure to hash the content could lead to it being forwarded or reposted on the service.

Third party datasets

A4.30 In the June 2025 Consultation, we noted the existence of third-party datasets, including those operated by the Global Internet Forum to Counter Terrorism (GIFCT) and Tech Against Terrorism (TAT). We said that at that stage, it was our understanding that there was currently not a "person with expertise" who met the requirements we had proposed should define an appropriate database, in particular because the databases made available by third parties were not limited to hashes of terrorism content as defined in the Act but also contained hashes of other terrorist and violent extremist material. We have changed our view on this point.

A4.31 We have not assessed GIFCT in detail for the purposes of this decision because it is not available to all service providers within the scope of this measure. This should not be taken to indicate that we consider GIFCT incapable of meeting our assessment criteria.

A4.32 We have undertaken a detailed review of TAT's Hash List API service and consider that it currently meets our criteria for an appropriate set of hashes. We therefore consider that at least one third party organisation exists that maintains an appropriate database and makes it available to service providers. Our fuller assessment of TAT's content hash list is set out in the next chapter of this annex, annex 4.5: Hash matching for terrorism content: assessment of Tech Against Terrorism content hash list.

Appropriate configuration of perceptual hash matching technology is necessary to ensure its effectiveness

A4.33 Service providers need to configure the perceptual hash matching technology to strike an appropriate balance between precision and recall.[111] This is essential for ensuring the accuracy and effectiveness of the technology and helping to mitigate against the risks of false positives. The Information Commission's Office (ICO) also supported this approach, endorsing the need to achieve an appropriate balance between precision and recall.[112]

A4.34 In the June 2025 Consultation, we proposed that service providers should take into account certain factors in striking this balance between precision and recall, including the risk of harm from terrorism content on the service, reflecting the service's risk assessment and any other information reasonably available to the provider; the proportion of content detected as a match by the technology that is a 'false positive'; and the effectiveness of the systems and processes used to identify false positives.[113] These principles aim to help ensure the accuracy and effectiveness of the technology and the way it is deployed overall on the service, and should therefore limit the risk of detection of content that is not terrorism content.

A4.35 What constitutes an appropriate balance depends on a range of factors, including the individual service context, the characteristics and prevalence of terrorism content on the service, and the downstream actions taken in response to detection. In addition, different implementations of perceptual hash matching may use different similarity measures, distance metrics or configuration parameters, meaning there is no single operating point that could be specified that would universally ensure accurate and effective detection of terrorism content. Further, we expect providers to review the performance of their perceptual hash matching systems over time, as the appropriate balance between precision and recall may change as content patterns, risks and system performance evolve. For all these reasons, it would be neither appropriate or desirable for Ofcom to prescribe the appropriate balance between precision and recall in our Codes of Practice.

A4.36 Providers should therefore determine the appropriate balance between precision and recall by reference to evidence from their testing, evaluation, and ongoing monitoring, recognising that the appropriate balance may change over time.

A4.37 A decision to optimise for precision may be appropriate, for example, where the provider has assessed its service as being medium risk of terrorism content and there is no evidence that the technology is failing to detect such content. This may particularly be the case where a relatively high proportion of detected content transpires to be false positives and the systems and processes in place to identify and correct false positives are less effective.

---

[111] Precision is the proportion of identified cases that are true positives. Recall is the proportion of true positive cases that are correctly identified.
[112] ICO response to the June 2025 Consultation, p.16.
[113] June 2025 Additional Safety Measures Consultation.

A4.38 Conversely, a decision to optimise for recall may be appropriate where the provider has assessed its service as being at high risk of terrorism content, including where there is evidence that known terrorism content is present on the service but remains undetected by the perceptual hash matching technology. This may be particularly relevant where a relatively low proportion of detected content transpires to be false positives and effective systems and processes are in place to identify and address false positives, including through secondary review of detected content.

A4.39 The measure is designed to allow flexibility, and this approach enables the measure to be adopted more widely. In practice, service providers have reputational and commercial incentives to encourage them to strike a reasonable balance in their implementation of the measure. We address the measure's impact on human rights in the section titled 'Rights impact' in chapter 10: Hash matching – Measure on perceptual hash matching for terrorism content.

Service providers should review the balance between precision and recall at least every six months

A4.40 In the July 2025 Consultation, we proposed that service providers should review the balance between precision and recall at least every six months, in order to secure that the balance is updated if appropriate. We have not received feedback to suggest that the six-month review period is disproportionate. We are making clear that service providers should review the balance between precision and recall at least every six months, or sooner where issues are identified. Service providers should not wait until the next scheduled review to address known issues. In particular, where a provider becomes aware of issues affecting the precision or recall of the technology, we expect it to take timely steps to assess and, where necessary, improve the technology's performance.

A4.41 We consider that this approach strikes an appropriate balance between ensuring the ongoing effectiveness of the technology and maintaining a proportionate burden on service providers. A review at least every six months ensures regular and structured oversight of performance, while the expectation that providers act sooner where issues arise provides flexibility and helps ensure that any deterioration in effectiveness is addressed without undue delay.

Written record

A4.42 We are also recommending that service providers maintain a written record of how this balance has been achieved. This should help to secure that the basis for decisions can be checked, so that their consequences can be understood and learned from.

Revision of terminology: replacing "visual images (whether or not combined with written material)" with "visual material"

A4.43 We have amended the terminology from "visual images (whether or not combined with written material)" to "visual material". This change simplifies and clarifies the wording, as "visual material" already encompasses images both with and without accompanying written content, making the text easier to read and understand.

A5. Assessment of Tech Against Terrorism's Hash List API service

Assessment of whether Tech Against Terrorism's Hash List API service meets the criteria for an "appropriate set of hashes"

Introduction

A5.1 In the June 2025 Additional Safety Measures Consultation (June 2025 Consultation), we provided an overview of our understanding of third-party databases and explained that at that point, it was our understanding that there was not currently a 'person with expertise' particularly because third-party databases were not limited to hashes of terrorism content as defined in the Online Safety Act 2023 (the Act) and also contained hashes of other terrorism and violent extremist material.[114]

A5.2 Following further review, including assessment of additional information relating to Tech Against Terrorism ('TAT'), specifically its Terrorist Content Analytics Platform (TCAP) hash list API Service, we have revised our position on this point.[115] We now consider Tech Against Terrorism to be one example of an organisation with expertise in the identification of terrorism content, and its hash list API service meets our criteria for an appropriate set of hashes (see Chapter 10, How this measure works). We outline our assessment of this provider below.

Background on Tech Against Terrorism

A5.3 Tech Against Terrorism was launched by the United Nations (UN) in 2016 and is a UK based independent non-profit organisation employing an interdisciplinary team of open-source intelligence analysts, data scientists and specialists in counter terrorism policy and human rights. It has worked with various online services across the tech sector and is supported by the UN. It has previously received funding for its work from various online services and a number of international governments including the UK, Canada, Australia and New Zealand.[116]

A5.4 Tech Against Terrorism's TCAP Content offers an API-based solution to online services. Tech Against Terrorism's TCAP Archive Database Hash List API Service allows online services to ingest cryptographic and perceptual hashes produced from known terrorism content in bulk for use in their content moderation processes.[117]

---

[114] Para 12.26 of the June 2025 Consultation.
[115] We refer to this as the "TCAP hash list", "Hash List API service", or "hash list" interchangeably throughout this chapter and annex.
[116] Tech Against Terrorism, About Tech Against Terrorism: Disrupting terrorist activity online.
[117] Tech Against Terrorism, Archive Database Hash List API; Terrorist Content Analytics Platform (TCAP), Tier 3: Designation.

Use of AI to detect and review content before it is added to a database

A5.5 TAT has informed us that it is developing and deploying a new AI capability to increase the scale at which terrorism content can be detected and reviewed online. We have not assessed this capability as part of our review.[118]

A5.6 We recognise that both service providers and third-party database providers may use AI to support the identification and review of content at scale. However, for the purposes of this measure, our assessment below relates only to TAT's Archive Database Hash List API Service, which contains content that has been reviewed by human reviewers.[119] We consider this database to meet our criteria.

A5.7 We have not assessed the effectiveness, accuracy, or suitability of TAT's new AI capability at this stage. TAT has confirmed that the human-reviewed database will continue to be available through its API, and that providers will be able to choose to use either or both datasets. TAT has also indicated that the human-reviewed database will continue to be updated, although growth is expected to be slower than for the AI-supported database.[120]

Assessment of the Hash List

A5.8 In the below section, we have considered the criteria of appropriate set of hashes against information we have received from Tech Against Terrorism.

Include hashes of terrorism content sourced by a person with expertise in identification of terrorism content

The database includes hashes of terrorism content

A5.9 We consider that for a database to meet this criterion, it must:

a) include content capable on its face of amounting to terrorism content; and
b) not include material volumes of content which is not content capable on its face of amounting to terrorism content.

A5.10 In that light, we have carefully considered the Inclusion Policy relating to TAT's Hash List API, the way it is applied in practice, and the degree to which filtering options are available to its users.

A5.11 TAT offers a Tiered Alert System Inclusion Policy. The TCAP Tiered Alert system is a prioritisation framework for alerting tech platforms in relation to terrorist and violent extremist content.[121] Where appropriate, TAT then hashes and archives relevant data into the hash list. The Inclusion Policy contains the following categories:

a) Tier 1 (Threat to Life): "content indicating a high threat to life";

b) Tier 2 (Crisis) (ongoing incident/historical incident) "content produced by terrorist attack perpetrators;

c) Tier 3 (Designation) "official propaganda by terrorist entities designated by selected democratic nation states and supranational institutions"; and

[118] Confirmed in TAT's written response dated 6 October 2026 – Questions 1, 2 and 3 relating to Artificial Intelligence.
[119] Confirmed in TAT's written response dated 6 October 2026 – Questions 1, 2 and 3 relating to Artificial Intelligence.
[120] Confirmed in TAT's written response dated 6 October 2026 – Questions 1, 2 and 3 relating to Artificial Intelligence.
[121] Tech Against Terrorism, *Inclusion Policy*. Terrorist Content Analytics Platform.

d) Tier 4 (Promotional) (unofficial media, instructional, inspirational) "material that directly supports designated terrorist organisations, glorifies terrorists or terrorist acts, or provides instructions for the purpose of terrorism".

A5.12 We recognise that these definitions do not all map exactly onto the definitions of UK terrorism content. However, we are now satisfied that the database, with filtering does, in practice, comprise content capable of amounting to UK terrorism content and that this is unlikely to change going forward for the following reasons:

a) TAT has explained in relation to Tier 1 that in recent years it has compiled threats to life reports manually and issued them directly to service providers, but has "rarely, if ever, collected TTL content for the purpose of archival retention".[122]
b) As to Tier 2, we consider that examples of historic crisis content may include content related to the Christchurch attack and the Buffalo attack. We recognise that such content may be reposted for the legitimate purposes of news reporting, academic analysis, or public discussion. However, it is still valuable for such content to be moderated as it may also be reposted by those who wish to inspire similar attacks by others. This highlights the importance of contextual assessment.
c) Within Tier 3 (Designation) we now understand that it is possible for service providers to filter the hashes so as to exclude hashes of content which relates to TAT's "designated terrorist entities" which are not proscribed in the UK.[123] [124] We consider that content which is 'the official propaganda of UK proscribed organisations' is likely to amount to multiple UK terror offences.
d) Tier 4 comprises 'unofficial media outlets'[125]; 'inspirational material'; and 'instructional material'. The detailed definitions of these generally map well to the UK terror offences, including s.1 Terrorism Act 2006 (encouragement of terrorism), s.2 Terrorism Act 2006 (dissemination of terrorist publications), s.58 Terrorism Act 2000 (collection of information likely to be of use to a terrorist); and the two offences s.54 of the Terrorism Act 2000 and s.6 of the Terrorism Act 2006 which, together, comprise the terrorist training offences. To the extent that an unofficial media outlet may relate to a non-UK proscribed organisation we are satisfied that this does not affect our overall analysis of the database. TAT has told us that currently all items tagged as an unofficial media outlet also carry the tag of a designated terrorist entity (and accordingly can be filtered). This is subject to change, but TAT has told us it prefers to assign content to tier

[122] TAT's informal response dated February 2026 (question 6) confirmed in RFI dated 2 September 2026 (Question 9).
[123] Confirmed in RFI dated 2 September 2026 – Question 1. The TCAP Content Hash List includes "designated terrorist entities" which are not currently proscribed in the UK. For a full list of UK proscribed organisations please see: Home Office, 2026, Proscribed terrorist groups or organisations. GOV.UK. Service providers can filter for UK proscribed organisations. We note that 'designation' in this context is TAT's term, which does not correspond with the definition of 'designation' in UK law (the National Security (State Threats) Act 2026). Designation under the National Security (State Threats) Act 2026, does not affect the definition of the priority terrorism offences under the Act.
[124] We understand that at present TAT does not treat Palestine Action as a 'designated terrorist entity'. Were it to choose to do so in future, we would expect service providers to apply UK definitions of proscribed organisations in determining whether or not content is illegal content.
[125] TAT has told us that it includes "content which bears the branding or insignia of the media outlet in question is sufficient to justify inclusion. This is because it is sufficiently certain that e.g. Al-Malahem operates exclusively in service of AQAP and likewise Al-Azaim in respect of ISKP." Source: TAT's informal response dated February 2026 (Question 12), confirmed in RFI dated 2 September 2026 (Question 9).

3, so we do not consider it likely that a material number of items will fall into this category in future. [126]

TAT's expertise in the identification of terrorism content

A5.13 As set out in Annex 4.4 (detail reasoning on effectiveness), a person with expertise in the identification of terrorism content may be, amongst other things, an expert third party database provider. We consider that as an interdisciplinary team of open-source intelligence analysts, data scientists and specialists in counter terrorism policy and human rights, funded by a number of international bodies, TAT is a person with appropriate expertise in identifying terrorism content.

A5.14 We consider that TAT has appropriate arrangements in place to identify and verify suspected terrorism content before hashes of the material are added to the database. All content collected for potential inclusion is reviewed against the inclusion policy by at least one intelligence analyst within 48 hours of collection before being added to the Hash List. [127]

Include an appropriate range of content amounting to terrorism (an offence under Schedule 5 of the Act)

A5.15 We consider that TAT's Inclusion policy covers a large number of the UK terrorism offences and captures content relating to a range of different terrorist ideologies.

A5.16 We consider that TAT has appropriate arrangements in place to identify a variety of suspected terrorism content. We understand it collects content for inclusion in the database from beacon channels [128] and through manual and automated processes. These include keyword searches, monitoring of terrorist and violent extremist channels and outlets, targeted investigations, and automated scrapers including AI tooling. [129]

Contain hashes of a significant number of original items of content

A5.17 The volume of items on the database is continually changing, but it contains hashes relating to more than 50 thousand items of content and as set out above we consider that the vast majority of these are items capable of amounting to terrorism content under the Act. [130]

Not plainly discriminate on the basis of protected characteristics (within the meaning of Part 2 of the Equality Act 2010)

A5.18 TAT's inclusion criteria suggest that that it does not discriminate on the basis of protected characteristic, since they contain no grounds for such discrimination. TAT has explained that the database includes content associated with a broad range of terrorist ideologies, including both Islamist and far-right terrorist organisations. [131]

---

[126] TAT's informal response dated February 2026 (question 13) confirmed in RFI dated 2 September 2026 (question 9) and follow up email dated 4 November 2026.
[127] Confirmed in RFI dated 2 September 2026 - Question 2G.
[128] Confirmed in RFI dated 2 September 2026 – Question 2. Beacon channels are platforms that terrorists, violent extremists and their supporters use to spread terrorist and violent extremist content to the widest possible audience. Official or promotional content is posted there, redirecting the audience to the platforms where the content is originally hosted. The beacon channel acts both as a central repository and as a signpost to where the content can be found.
[129] Confirmed in RFI dated 2 September 2026 - Question 3.
[130] Confirmed in RFI dated 2 September 2026 - Question 10 and Appendix 1; Confirmed in TAT's written response dated 6 October 2026 - Question 2 (information request follow up).
[131] Tech Against Terrorism, Terrorist Content Analytics Platform (TCAP), Tier 3: Designation.

Be proactively updated with reasonable regularity

A5.19 TAT has confirmed that it continuously identifies new content through analyst-led investigations and automated collection tools, which operate on daily, weekly and monthly schedules. Updates, including additions and removals, are then made available to users through the API.[132]

Be proactively updated to remove incorrect hashes, where appropriate. The service provider should take reasonable steps to secure the removal of a hash from the database where the provider has determined that the underlying content is not capable of being terrorism content

A5.20 TAT has told us it has processes in place to support the proactive removal of incorrect hashes. In particular, if an entity previously within scope were no longer considered eligible for inclusion, TAT indicated that it would identify and withdraw the associated underlying content from the Archive, which in turn would remove the corresponding hashes from the hash list.[133] TAT also stated that internal administrators can remove content where there are no longer grounds for inclusion and that users can report potential false positives through its Feedback API for review by TAT analysts.[134]

A5.21 TAT reviews the nature of beacon channels to confirm their terrorist nature at least monthly.[135]

A5.22 Accordingly, TAT's evidence suggests that mechanisms exist to secure the removal of hashes where the underlying content is no longer considered terrorism content.

Be secured from unauthorised access or interference (whether by persons who work for that person or are providing a service to that person, or any other person)

A5.23 TAT has told us that access is restricted to authenticated current TAT staff, data is encrypted both in transit and at rest, and the platform is subject to regular penetration testing to assess the effectiveness of security controls.[136]

Dataset conclusion

A5.24 Having considered the information provided by TAT, we are satisfied that the Hash List meets the criteria for an appropriate set of hashes for the purposes of this measure.

[132] Confirmed in RFI dated 2 September 2026 - Question 5a.
[133] Confirmed in RFI dated 2 September 2026 - Question 6.
[134] Confirmed in RFI dated 2 September 2026 - Question 6.
[135] Confirmed in RFI dated 2 September 2026 -Question 2d.
[136] Confirmed in RFI dated 2 September 2026 - Question 11.