Proactive Technology Draft Guidance

Additional Safety Measures – Annex 6

Consultation

Published 30 June 2025
Closing date for responses: 20 October 2025

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

Contents

Section

1. Overview

What this guidance covers

This is our guidance to assist providers of regulated user-to-user services in implementing proactive technology for the purpose of detecting illegal content and content harmful to children.

This guidance is applicable for the purpose of understanding how proactive technology should be implemented where necessary for the recommended measures set out in Ofcom's Illegal Harms User-to-user Code and Protection of Children User-to-user Code.

Navigating the guidance

1.1 The guidance consists of the following sections:

- Section 2: Introduction
- Section 3: The proactive technology criteria
- Section 4: Understanding the assessment process
- Section 5: Understanding and applying the proactive technology criteria

2. Introduction

Background to the draft guidance

Illegal Content Codes of Practice and Protection of Children Code of Practice

2.1 In our December 2024 Statement on Protecting People from Illegal Harms Online (December 2024 Statement) and the April 2025 Statement on Protecting Children from Harms Online (April 2025 Statement) we recommended that services review, assess, and swiftly take down – or take action on – illegal content, illegal content proxy,[1] content harmful to children and/or content harmful to children proxy[2] when they become aware of it and where it is technically feasible to do so (Measures ICU C1 and C2 and Measures PCU C1 and C2).[3]

2.2 In our December 2024 Statement we also recommended that providers should use hash-matching and URL detection technology to detect and remove child sexual abuse material (CSAM) and URLs identified as hosting CSAM (Measures ICU C9 and ICU C10).[4] These technologies are not in-scope of this guidance.

2.3 Further to consideration of responses to the November 2023 Consultation on Protecting People from Illegal Harms Online (November 2023 Consultation) and the May 2024 Consultation on Protecting Children from Harms Online (May 2024 Consultation), we are now proposing additional measures about the use of proactive technology.[5]

[1] Illegal content proxy: In the Illegal Content User-to-user Codes, we defined "illegal content proxy" as content that a provider determines to be in breach of its terms of service, where: the provider has reason to suspect that the content may be illegal content; and the provider is satisfied that its terms of service prohibit the type of illegal content which it had reason to suspect existed.

[2] Content that is harmful to children proxy: In Codes, we define "content that is harmful to children proxy" as primary priority content (PPC) proxy, priority content (PC) proxy or non-designated content (NDC) proxy". This is content that a provider determines to be in breach of its terms of service, where: a) the provider had reason to suspect that the content may be relevant PPC, PC and/or NDC; and b) the provider is satisfied that its terms of service prohibit the type of relevant priority content which it had reason to suspect existed.

[3] December 2024 Statement on Protecting People from Illegal Harms Online (December 2024 Statement). Volume 2 Service Design and User Choice. Chapter 2. [accessed 13 June 2023] and April 2025 Statement on Protecting Children from Harm Online (April 2025 Statement). Volume 4 What should services do to mitigate the risks of online harms to children. Chapter 1. [accessed 13 June 2025].

[4] December 2024 Statement. Volume 2 Service Design and User Choice. Chapter 4. [accessed 13 June 2025].

[5] See Chapter 6: Proactive Technology, section 'What risk does the use of proactive technology seek to address?'

The proactive technology measures

Table 2.1: Proposed measures and who should implement this

Number	Proposed measure	Who should implement this
ICU C11	Assessing proactive technology for use to detect or support the detection of target illegal content. [6]	• large user-to-user services that are medium or high risk for at least one relevant harm • user-to-user services with more than 700,000 monthly UK users that are high risk, for at least one relevant harm • user-to-user services that are file-storage and file-sharing services which identify a high risk of image-based CSAM, regardless of size • All user-to-user services which identify a high risk of grooming
PCU C9[7]	Assessing proactive technology for use to detect or support the detection of target content harmful to children.[8]	
ICU C12	Assessing existing proactive technology for use to detect or support the detection of target illegal content.	
PCU C10[9]	Assessing existing proactive technology for use to detect or support the detection of target content harmful to children.	

2.4 We are recommending these measures apply for the following harms:

- Illegal harms: child sexual exploitation and abuse (CSEA), encouraging or assisting suicide (suicide), and fraud and other financial services offences (fraud); and
- Content harmful to children: primary priority content (PPC), which includes suicide, self-harm, eating disorder content and pornographic content.

ICU C11 and PCU C9

2.5 First (see section 'Sourcing proactive technology: Step 1'), when assessing whether proactive technology can be deployed, providers should:

- identify whether proactive technology that detects or supports the detection of target illegal content and/or content harmful to children is available;

[6] **Target illegal content** means **relevant content** that a) amounts to an offence in relation to the **relevant harms**; or b) is **illegal content proxy**, where the provider is satisfied that its ***terms of service*** prohibit **the relevant harms**.

[7] This measure is limited to services likely to be accessed by children.

[8] **Target content harmful to children** means **relevant content** that either: a) relates to one of the following: (i) where the service is a large service, the specific kinds of primary priority content for which the service is at medium or high risk, to the extent they are relevant primary priority content; or (ii) where the service is not a large service, the specific kinds of primary priority content for which the service is at high risk, to the extent they are relevant primary priority content; or b) is **content that is harmful to children proxy**, where the provider is satisfied that its terms of service prohibit the specific kinds of relevant primary priority content.

[9] This measure is limited to services likely to be accessed by children.

- identify whether it is technically feasible to implement that proactive technology on their service; and
- assess whether the proactive technology meets the proactive technology criteria.

2.6 Second (see section 'Sourcing proactive technology: Step 2'), where proactive technology is available, is technically feasible to implement on their service, and meets the proactive technology criteria, the provider should procure and deploy it. Once the technology is deployed on the service, a provider will be considered to have complied with the measure.

2.7 The proposed measure will only apply where it is technically feasible for a service to implement proactive technology. We do not consider that it would be technically infeasible to implement proactive technology merely because to do so would require some changes to be made to the design and/or operation of the service.

2.8 However, the proposed measure will not apply to providers for whom it is not technically feasible to analyse user-generated content present or disseminated on the service to assess whether it is content of a particular kind, particularly where such changes as would need to be made to enable this would materially compromise the security of the service.

2.9 Where providers conclude that it is not technically feasible to implement the proactive technology on their service, they should keep a record of this decision together with reasons why. For details, see 'Where providers cannot deploy proactive technology'.

Providers who choose to develop proactive technology in-house.

2.10 Some providers may consider it appropriate to develop proactive technology in-house to detect the relevant harms proposed in this measure. Providers may take this approach provided the in-house technology meets the proactive technology criteria.

2.11 Providers who decide to develop proactive technology in-house need not conduct the initial assessment (see paragraph 2.5) but the proactive technology criteria should still be applied when developing and deploying these solutions.

2.12 When deciding whether to develop proactive technology in-house, providers should initially consider the following factors:

- **Internal capability and expertise**: Providers should consider whether they have access to the necessary technical expertise to design, build, evaluate and maintain the proactive technology and/or combination of technologies over time.
- **Governance and accountability**: Providers should have clearly defined accountability and ownership for the development, performance, and ongoing compliance of the proactive technology or combination of technologies.
- **Delivery timelines**: Providers should consider whether they have a clear and realistic development timeline which includes consideration of risks to delivery.
- **Fallback options:** Providers should have a plan in place to mitigate the risk of failure or delay, including the use of interim tools if necessary.

2.13 While this guidance is focused on providers who source proactive technology, the proactive technology criteria should still be applied for those who choose to develop proactive technology. For further information on developing proactive technology see section: 'Developing proactive technology'.

ICU C12 and PCU C10

2.14 We propose ICU C12 and PCU C10 apply to providers who

- have already deployed proactive technology to detect target illegal content and/or content harmful to children before the implementation of ICU C11 and PCU C9; and/or
- who have deployed proactive technology to comply with ICU C11 and PCU C9.

2.15 In summary, we propose that such providers should:

- Assess the proactive technology that they have already deployed to detect or support detection of target illegal and/or content harmful to children against the proactive technology criteria;
- if necessary, take steps within a reasonable time to ensure that the proactive technology meets all the proactive technology criteria or where the proactive technology cannot be changed to meet the criteria, to source new proactive technology; and
- record the outcome of their assessment.

2.16 For further information see 'Measure ICU C12 and PCU C10: Assessing existing proactive technology for use to detect or support the detection of target illegal content and/or content harmful to children.'

Purpose of this draft guidance

2.17 This guidance aims to help service providers implement the recommended measures that relate to the deployment of proactive technology for the detection of illegal content and/or content harmful to children. The guidance provides additional detail and practical examples to help service providers assess whether a proactive technology meets the proactive technology criteria. Where it does, the guidance also supports providers to understand steps they can take to help ensure the proactive technology continues to meet the criteria once deployed and throughout operational use.

2.18 The guidance is intended to support all providers in scope of ICU C11, PCU C9 ICU C12 and PCU C10, whether sourcing from a third-party supplier, developing proactive technology in-house, or reviewing an existing deployment of a proactive technology.

2.19 Many of the activities described (particularly the illustrative examples) are applicable to all scenarios. However, the primary focus of the guidance is to support providers who are assessing proactive technology sourced from third party suppliers to assist in making an assessment against the criteria prior to deployment and, where the criteria are met, to support continued evaluation and compliance throughout the technology's lifecycle. We anticipate that providers capable of developing proactive technology are likely to be familiar with the processes involved or will already have established procedures in place. Regardless of the approach taken, all providers in scope of the measures should ensure that any proactive technology deployed meets the proactive technology criteria.

3. The proactive technology criteria

3.1 The Online Safety Act 2023 (the Act) requires Ofcom to have regard to the degree of accuracy, effectiveness and lack of bias achieved by a technology in deciding whether to include it as a proactive technology measure in a Code of Practice or in a confirmation decision.[10] This helps minimise disproportionate impacts on privacy and freedom of expression (for example, instances of content being wrongly removed due to proactive technology).

3.2 The Act also allows Ofcom to set out principles in a Code of Practice designed to ensure that proactive technology or its use is (as far as possible) accurate, effective and free of bias.[11]

3.3 The concepts of accuracy, effectiveness, and lack of bias are not defined within the Act. For the purpose of developing this measure, we have broadly defined these concepts as follows:

- **Accuracy:** the technical correctness and reliability of the proactive technology as assessed during development, testing, and evaluation, including the extent to which errors are minimised through evidence-based methodologies and appropriate performance metrics.
- **Effectiveness:** the utility and ability of the proactive technology to achieve intended outcomes, evaluated by assessing its real-world impact, alignment with specific goals, and operational relevance for the intended use case.
- **Lack of bias:** the extent to which the proactive technology avoids discrimination and negative impacts on equity and fairness, including with respect to the treatment of different users and the handling of different types of content or other inputs.

3.4 For further explanation please see ‘What is Proactive Technology’.

The criteria

3.5 In having regard to accuracy, effectiveness and lack of bias, we have developed a set of criteria that we propose providers should consider when assessing and deploying proactive technology.

3.6 Where proactive technology can be deployed for the purpose of detecting target illegal and/or content harmful to children in such a way that the technology meets all the proactive technology criteria, we consider that the technology is sufficiently accurate, effective, and lacking bias such that it is proportionate to recommend its implementation.

3.7 The proactive technology criteria are as follows:

[10] Schedule 4 to the Act, paragraph 13(6).
[11] Schedule 4 to the Act, paragraph 13(6)(b).

Table 3.1: Proactive technology criteria

	Proactive technology criteria
Use of high-quality data	The proactive technology has been developed and tested using high-quality[12] datasets appropriate to and reflecting a broad range of inputs relevant to the harm it is intended to detect (as identified in the services' risk assessment.
Addressing biases	Potential biases have been identified and addressed during the design and development process, and risks are appropriately managed and addressed throughout the proactive technology's lifecycle.
Evaluating performance	The proactive technology has been evaluated using appropriate performance metrics and configured so that its performance strikes an appropriate balance between precision[13] and recall.[14] [15]
Safeguards against misuse and exploitation	Safeguards are in place to identify and appropriately manage security threats and risks of exploitation and misuse, including through the use of access restriction[16] and system integrity protections.
Contextual testing and evaluation	The proactive technology's performance has been evaluated in real-world use cases relevant to the provider's content (having regard to the risk of harm to individuals identified in the providers' risk assessment(s)) and the results indicate it correctly detects the harm it is intended to detect. This includes testing for scalability, handling of different media types (where relevant), and whether the proactive technology's performance could be improved by layering with complementary approaches or (in the case of existing deployments) by updating to a more current version.
Maintenance and ongoing monitoring	Mechanisms are in place to monitor and maintain the proactive technology's effectiveness over time, including processes for regular review and iterative adjustments to respond to emerging circumvention techniques, biases or new content types.

[12] "High quality" refers to data that is accurate; legally and ethically sourced; well-labelled (where appropriate); representative of the harm and context the proactive technology is intended to address; and sufficiently diverse to support meaningful evaluation, including (where relevant) to test the proactive technology's ability to detect content it has not previously encountered. The degree to which a dataset may be considered "high-quality" will vary based on the harm being addressed and the context in which the proactive technology is used. However, indicators that a dataset might not be considered "high quality" would include (for example) unclear sourcing, evidence of poor or inconsistent labelling, or limited representation of relevant harms or input types.

[13] Precision is the proportion of identified cases that are true positives.

[14] Recall is the proportion of true positive cases that are correctly identified.

[15] Please see 'C: Evaluating Performance: What should providers do' for certain factors providers should consider related to the testing, configuration, deployment and ongoing monitoring of the proactive technology.

[16] "Access restrictions" refers to limitations placed on interaction with system operations or data, based on the roles or responsibilities assigned to any entity (including individuals, devices, or systems such as software components, applications, and automated processes), whether internal and external. This aligns with principles on controlling access, as outlined in the UK Cyber Assessment Framework, and supported by international standards such as ISO/IEC27001 and NIST SP-800-53.

	Proactive technology criteria
Human review	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 judgement and accountability.[17]
Incorporating feedback	Feedback mechanisms are in place to maintain or improve performance over time. This includes updating the proactive technology with diverse and up-to-date datasets to reflect evolving trends or emerging types of illegal content and/or content harmful to children and/or integrating ongoing feedback from users and individuals working in content moderation[18] into its development, while managing the risk of introducing additional bias.

3.8 In Section 4, we set out the process that providers should follow to determine whether proactive technology meets the proactive technology criteria.

3.9 In Section 5, we explore the criteria in detail, outlining our rationale and illustrative examples of how each criteria may be assessed.

[17] Please see 'G: Human Review: What should providers do' for certain factors providers should consider related to the testing, configuration, deployment and ongoing monitoring of the proactive technology.

[18] For an explanation of 'individuals working in content moderation' see Illegal Harms Statement Volume 2: Service design and user choice pp. 79-80. [accessed 12 June 2025]

4. Understanding the assessment process

4.1 This section outlines the process providers should follow to assess whether proactive technology meets the proactive technology criteria.

4.2 For the purposes of this guidance, we have grouped activities recommended as part of both measures, into three high-level stages:

i) initial assessment;
ii) testing and configuration; and
iii) deployment and ongoing monitoring.

4.3 While these stages do not map precisely onto a formal technology lifecycle, they are intended to capture typical decision points and activities that we expect providers will undertake when assessing, deploying and/or reviewing proactive technology.

4.4 Although we have used this structure to reflect how providers might engage with the criteria in practice, it is not intended to be prescriptive. Providers have flexibility to take the approach that best suits their service context, operating model and the nature of the proactive technology being assessed. Whatever approach is taken, as part of the assessment process providers should consider how each criterion can be met across the full lifecycle of the proactive technology, including how compliance will be maintained after deployment.

4.5 We note that providers should comply with the law, including criminal law, when implementing proactive technology that meets the proactive technology criteria. For example, we note that there are priority offences relating to making, showing, distributing or possessing an indecent image or film of a child; an offence of possession of a prohibited image of a child; and an offence of possession of a paedophile manual. Providers should ensure they are not committing these offences while implementing this measure, for example, in the course of compiling datasets on which proactive technology is trained.

4.6 In the following paragraphs we set out the steps providers should follow when assessing whether proactive technology designed to detect or support the detection of target illegal content and/or content harmful to children meets the proactive technology criteria.

4.7 Further details of how providers might apply the proactive technology criteria when assessing a proactive technology, including illustrative examples of practical steps, are provided in Section 5: Understanding and applying the proactive technology criteria.

4.8 While not exhaustive, we provide examples of activities that a provider seeking to source an existing proactive technology may undertake at each of these stages.

Initial assessment:

4.9 Providers may, as part of their initial assessment, conduct market research or engage with potential supplier(s). As part of this assessment, they may review documentation and evidence to assess the proactive technology's capability of meeting the criteria, including planning for human review.

4.10 Some activities (such as full testing or implementation of safeguards) may not be possible at the initial assessment stage. However, we anticipate that providers should be able to gather sufficient evidence (for example, through market research) to form a reasonable judgement about whether the proactive technology is likely to be capable of meeting the criteria in full once later assessment stages are reached. For example, if a proactive technology supports output logging and integration with moderation workflows, and the provider has, or clearly can, develop a plan for human oversight after deployment, this would generally be sufficient to support a decision to proceed to the next stage of assessment.

4.11 Where this initial assessment is carried out thoroughly and based on clear evidence, it should provide a strong indication that the proactive technology can meet the criteria, reducing the risk of investing time and resource in a solution that is ultimately deemed not to meet the requirements of these measures.

Testing and configuration

4.12 Providers should validate the performance of proactive technology in the context of the service. This may include trial deployments or sandbox testing, configuring settings such as detection thresholds, and confirming the proactive technology can be adapted to reflect service-specific risks and input types. This stage may also include more detailed scoping and testing of the human review process, including workflows and the volume and frequency of outputs to be reviewed.

4.13 During this stage, we would expect a provider to:

- Understand how the proactive technology functions in a variety of real-world scenarios relevant to the service on which they will be deployed.
- Ensure that the proactive technology is configured so that its performance strikes an appropriate balance between precision and recall.
- Establish the appropriate proportion of outputs from the proactive technology to be subjected to human review.

Deployment and ongoing monitoring:

4.14 Once deployed, providers should ensure the proactive technology continues to meet the criteria over time through regular monitoring, updating and refinement. This includes addressing changes in patterns of harm, emerging risks (such as new circumvention techniques) and maintaining appropriate oversight through human review.

4.15 During this stage, we would expect a provider to:

- Ensure that any changes to the service, user behaviour, manifestations of harm, or the technology itself do not undermine the effectiveness of the proactive technology.
- Refine the proactive technology as or when necessary, so that its performance continues to strike an appropriate balance between precision and recall.

5. Understanding and applying the proactive technology criteria

5.1 This section sets out each of the proactive technology criteria in detail. For each of the proactive technology criteria, we explain:

- why it is important;
- what providers should do in order to assess proactive technology against the criteria and the type of questions providers might consider; and
- illustrative examples of practical steps that may be taken to meet the criteria.

5.2 We know many providers may already conduct similar exercises when sourcing or developing proactive technology to ensure that it meets their service's requirements.

5.3 The practical examples provided for each of the criteria are not exhaustive, however, they may help providers determine whether their proactive technology(ies) meets the proactive technology criteria. They have been included to support providers in their understanding of the types of activities that may be relevant at different stages of the assessment process and in different scenarios. We also acknowledge that providers may conduct different tests to achieve the same standard of confidence that the proactive technology criteria have been met.

5.4 The practical examples are relevant to all providers in scope of both ICU C11 and PCU C9 (whether sourcing proactive technology from a third party or developing it) and ICU C12 and PCU C10. However, the specific action a provider should take may differ depending on the context. For example:

- Providers who **source** proactive technology from a third party are likely to focus on gathering evidence from the supplier and verifying it through testing and monitoring.
- Providers who choose to **develop** proactive technology will typically carry out all activities directly as part of their design and implementation process.
- Providers **reviewing an existing deployment** of proactive technology (ICU C12 and PCU C10) may interpret the illustrative examples as prompts for assessing whether the proactive technology meets or continues to meet the criteria. This may involve reviewing historical documentation, performance data or supplier literature. Where gaps are identified or one or more of the criteria are not being met, providers should take appropriate action, such as updating, reconfiguring or layering/augmenting the proactive technology with other technologies to ensure the relevant criteria are met.

5.5 In considering how to perform their assessment, we expect providers to interpret and apply the criteria proportionately based on the context of their individual service, the nature of the harm, and the characteristics of the proactive technology under consideration.

5.6 This guidance may act as the baseline for assessing whether any proactive technology meets the criteria and is thus sufficiently accurate, effective and lacking in bias to be recommended for its implementation.

A: Use of high-quality data

What criteria must be met?

5.7 The proactive technology has been developed and tested using high-quality datasets appropriate to and reflecting a broad range of inputs relevant to the harm it is intended to detect (as identified in the service's risk assessment(s)).

Why is this important?

5.8 If the datasets do not reflect the diversity of inputs the proactive technology will encounter, it may fail to effectively detect harmful content or behaviour; or produce outputs which have discriminatory or negative impacts for different users.

What should providers do?

5.9 Assess whether the datasets used (for training, testing or operation) are sufficiently high-quality and appropriate to the harm and service context. Identify available mitigations which could be applied to address gaps or limitations.

What type of questions might providers consider?

- Has the proactive technology been tested on relevant types of target illegal content and/or content harmful to children?
- Has it been developed or tested using diverse datasets (for example, different types of content or other inputs and across user groups)?
- What is the origin of the data?

Illustrative activities

Table 5.1: Three stages and illustrative activity for criteria A: 'Use of high-quality data'

Stage	Illustrative activity
Initial assessment	• Identify the origin and composition of datasets used to develop or train the technology • Evaluate whether datasets align with the harm and input types that the proactive technology is intended to detect • Check documentation for diversity of sources, and (where relevant) labelling accuracy, and annotation standards
Testing and configuration	• Test system performance using representative datasets reflective of real service inputs • Assess consistency of detection across content formats and user groups • Modify dataset components or retrain using more representative samples if required
Deployment and ongoing monitoring	• Monitor for drops in detection quality due to unseen inputs or changing trends • Expand datasets based on missed cases or user-reported issues • Keep documentation of dataset updates and rationale

B: Addressing biases

What criteria must be met?

5.10 Potential biases have been identified and addressed during the design and development process, and risks are appropriately managed and addressed throughout the proactive technology's lifecycle.

Why is this important?

5.11 Bias introduced during design, development or in operational use can lead to discriminatory or unequal treatment of different user groups, types of content or other inputs if not addressed.

What should providers do?

5.12 Understand how bias has been or will be identified and addressed during development, testing and deployment. Consider how decisions made across the lifecycle of the proactive technology, such as in data selection, data labelling, rule design, or threshold setting (as relevant), may contribute to bias or unintended outcomes. Identify steps to monitor and mitigate any disparities.

What type of questions might providers consider?

- How has the proactive technology been tested across different user groups?
- How often is the proactive technology audited to identify and mitigate bias?
- If bias is detected, what steps are taken to mitigate this and do these steps adequately address bias?
- Is the proactive technology's decision-making process transparent?

Illustrative activities

Table 5.2: Three stages and illustrative activity for criteria B: 'Addressing biases'

Stage	Illustrative activity
Initial assessment	• Review evidence of bias evaluation from developer and/or vendor • Review the extent to which inputs and outputs are evaluated across different user demographics, types of content or other inputs • Consider service-specific bias risks and check if these have been or can be addressed
Testing and configuration	• Conduct bias testing on outputs of the proactive technology using representative samples of service-specific data • Identify disparities across relevant groups • Provide feedback to supplier and document mitigation steps
Deployment and ongoing monitoring	• Monitor outputs for new or emerging bias using user reports, sampling or audit tools • Schedule periodic bias reviews following updates or service changes • Refine or adjust review processes if patterns of unfair treatment are identified

C: Evaluating performance

What criteria must be met?

5.13 The proactive technology has been evaluated using appropriate performance metrics and configured so that its performance strikes an appropriate balance between precision and recall.

Why is this important?

5.14 Proactive technology used for detection of harmful content involves making trade-offs between false positives and false negatives. Understanding and managing those trade-offs is essential to ensure the proactive technology performs proportionately, balancing the risk of over-removal of legitimate content with failure to effectively detect harm.

What should providers do?

5.15 Understand how performance of the tool is measured and calibrated and whether the metrics used are appropriate to the proactive technology under consideration.

5.16 Assess whether the thresholds and settings are appropriate for the service context and risk the proactive technology seeks to address.

5.17 Ensure the proactive technology is configured so that its performance strikes an appropriate balance between precision and recall. In doing so, providers should ensure that the following matters are taken into account:

- the service's risk of relevant harm(s) proposed as part of this measure, reflecting the risk assessment of the services and any information reasonably available to the provider about the prevalence of target illegal content and/or harmful content on the service;
- the proportion of detected content that is a false positive;
- the effectiveness of the systems and/or processes used to identify false positives; and
- in connection with image-based CSAM and CSAM URLs, the importance of minimising the reporting of false positives to the National Crime Agency (NCA) or a foreign agency.

5.18 What constitutes an appropriate balance between precision and recall will depend on the nature of the relevant harm, the level of risk identified and the service context. For example, in some cases a provider might optimise for recall to maximise the quantity of content detected and apply additional safeguards, such as use of complementary tools or increased levels of human review, to address false positives. In other cases, higher precision may be more appropriate, for example, to reduce the risk of adverse impacts on user rights.

5.19 However, in circumstances where false positives are consistently high and cannot be meaningfully reduced or mitigated, particularly where this may have a significant adverse impact on user rights, providers may conclude that the proactive technology is incapable of meeting the criteria.

What type of questions might providers consider?

- What performance metrics (such as precision and recall) have been used to evaluate the technology?
- How have the trade-offs between false positives and false negatives been considered and managed?

- Does the proactive technology support configuration or tuning (such as threshold adjustment) to manage those trade-offs?
- Is the current configuration appropriate to the level of harm being addressed, the context of the service and its user base?

Illustrative activities

Table 5.3: Three stages and illustrative activity for criteria C: 'Evaluating performance'

Stage	Illustrative activity
Initial assessment	• Review performance benchmarks and understand which metrics (such as precision, recall, F1 score) are used • Assess whether the technology supports configurable thresholds • Review prior test results or independent evaluations and assess whether claimed performance aligns with service's risk thresholds
Testing and configuration	• Configure thresholds appropriate to service's risk profile and appropriately balance between false negatives and false positives • Validate performance in service-specific scenarios • Record configuration settings and rationale
Deployment and ongoing monitoring	• Continuously monitor performance metrics • Review performance logs regularly • Adjust settings if performance degrades or service conditions change

D: Safeguards against misuse and exploitation

What criteria must be met?

5.20 Safeguards are in place to identify and appropriately manage security threats and risks of exploitation and misuse, including through the use of access restrictions and system integrity protections.

Why is this important?

5.21 Proactive technology for content detection is susceptible to adversarial exploitation and misuse, potentially leading to compromised reliability. Safeguards such as input validation, evasion detection and configuration constraints help ensure the proactive technology functions as intended and maintains performance. Safeguards such as access restrictions and mechanisms to ensure system and data integrity are also required to address security risks and support reliable operation throughout the lifecycle of the proactive technology.

What should providers do?

5.22 Identify what safeguards are in place to prevent exploitation or misuse, including adversarial resistance mechanisms and fail-safes.

5.23 Understand the monitoring and alerting systems in place to identify unusual or unexpected behaviour.

5.24 Evaluate incident response and escalation protocols.

What type of questions might providers consider?

- How does the proactive technology detect potential attempts to manipulate or circumvent its operation?
- What fallback mechanisms exist for handling system failures or uncertainty? Are safeguards regularly tested and updated?
- What security measures are in place to prevent unauthorised access to the proactive technology or its underlying data throughout development, testing and deployment?

Illustrative activities

Table 5.4: Three stages and illustrative activity for criteria D: 'Safeguarding against misuse and exploitation'

Stage	Illustrative activity
Initial assessment	• Identify built-in safeguard features (such as access restrictions, output logging or protections against unauthorised modification) to mitigate risks of misuse, exploitation or security compromise • Evaluate how the technology handles ambiguous or adversarial inputs • Assess likelihood of circumvention in the service-specific use case/context
Testing and configuration	• Simulate misuse scenarios (such as obfuscation, evasion techniques) • Implement controls such as rate limiting or alerting on unexpected use • Log outcomes of misuse testing
Deployment and ongoing monitoring	• Regularly audit user and system behaviour for misuse indicators • Work with supplier to patch vulnerabilities or adjust safeguard mechanisms as required

E: Contextual testing and evaluation

What criteria must be met?

5.25 The proactive technology's performance has been evaluated in real-world use cases relevant to the service provider's content (having regard to the risk of harm to individuals identified in the providers' risk assessment(s)) and the results indicate it correctly detects the harm it is intended to detect. This includes testing for scalability, handling of different media types (where relevant), and whether the technology's performance could be improved by layering with complementary approaches or (in the case of existing deployments) by updating to a more current version.

Why is this important?

5.26 Pre-deployment evaluation in real-world or closely simulated conditions is essential to verify the proactive technology's ability to operate effectively in the context in which it will be deployed. This includes ability to operate at scale, handle relevant types of target illegal content and/or content harmful to children, and respond to actual usage patterns.

What should providers do?

5.27 Conduct testing in conditions that reflect actual service use. This should include assessing how the proactive technology performs across relevant types of target illegal content and/or content harmful to children or other inputs and evaluating and mitigating any observed degradation in performance.

What type of questions might providers consider?

- Can the proactive technology be tested in an environment that reflects our service context?
- Does performance vary across different content types, users and/or under high load?
- Can necessary adaptations or reconfigurations be implemented to mitigate identified issues, including by the addition of complementary technologies?

Illustrative activities

Table 5.5: Three stages and illustrative activity for criteria E: 'Contextual testing and evaluation'

Stage	Illustrative activity
Initial assessment	• Confirm whether the technology has been tested in a comparable service or content environment • Assess the match between detection capabilities and the types of content or other inputs on the service • Assess whether the proactive technology will require layering or adaptation to suit service context/use case
Testing and configuration	• Conduct trials using real or synthetic service data • Test scalability under high load and performance across different input types • Explore combining with complementary technologies for layered coverage to enhance performance
Deployment and ongoing monitoring	• Evaluate real-world performance and error patterns • Identify scenarios where performance may be degraded • Adjust configuration/augment with additional solutions as needed

F: Maintenance and ongoing monitoring

What criteria must be met?

5.28 Mechanisms are in place to monitor and maintain the proactive technology's effectiveness over time, including processes for regular review and iterative adjustments to respond to emerging circumvention techniques, biases or new content types.

Why is this important?

5.29 Ongoing monitoring and adjustment to the proactive technology are essential to maintain effectiveness of the system overtime and prevent degradation of performance as content evolves, new harms emerge and/or user behaviour changes.

What should providers do?

5.30 Establish monitoring processes to track performance of the proactive technology and identify and respond to emerging risks.

5.31 Update the proactive technology as necessary to respond to identified changes in performance.

What type of questions might providers consider?

- How is ongoing performance monitored and assessed?
- What triggers a review or an update of the proactive technology?
- How is, or can, the proactive technology be adapted to address newly identified risks?

Illustrative examples

Table 5.6: Three stages and illustrative activity for criteria F: 'Maintenance and ongoing monitoring'

Stage	Illustrative activity
Initial assessment	• Check whether the system supports logging and audit trails • Confirm that monitoring frameworks can be integrated • Assess availability/suitability of update mechanisms
Testing and configuration	• Establish monitoring baselines and thresholds • Simulate changes in user behaviour, types of content or other inputs and test system adaptation • Configure monitoring tools to integrate with internal systems
Deployment and ongoing monitoring	• Regularly track metrics (such as precision, recall, and volume metrics) • Implement process for engaging with supplier for retraining or adjustment of the proactive technology based on observed drift or circumvention • Log updates and regularly review effectiveness

G: Human review

What criteria must be met?

5.32 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 judgement and accountability.

Why is this important?

5.33 Human oversight is a well-recognised safeguard in the responsible use of automated systems, to assist in ensuring the outputs are accurate, fair and contextually appropriate. Human review supports accountability, protects user rights and enables corrective action where detection errors or edge cases arise. When outputs are explainable (meaning it is possible to understand how and why they were generated) reviewers are better able to assess and make informed decisions.

What should providers do?

5.34 Ensure human review is incorporated into the development and testing process to assess early outputs and inform adjustments to the proactive technology.

5.35 Ensure an appropriate proportion of outputs are subject to human review post-deployment. Providers have flexibility to decide what proportion it is appropriate to review, however in so doing, the following factors should be taken into account:

- 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;
- the principle that content with a higher likelihood of being a false positive should be prioritised for review; and
- in the case of image-based CSAM or CSAM URLs, the importance of minimising the reporting of false positives to the NCA or a foreign agency.

What type of questions might providers consider?

- What role does human review play in testing and refinement of the proactive technology prior to deployment?
- What proportion of outputs will be reviewed during deployment and how will this be determined?

Illustrative activities

Table 5.7: Three stages and illustrative activity for criteria G: 'Human review'

Stage	Illustrative activity
Initial assessment	• Determine whether outputs are accessible for human review • Assess the ability to sample or audit detection results • Plan human review scope and escalation thresholds
Testing and configuration	• Trial human review workflows and adjust as required • Evaluate consistency between human and system decisions • Integrate feedback from human review into system adjustments
Deployment and ongoing monitoring	• Maintain regular human review of appropriate proportion of outputs • Document intervention/overturn rates and review escalations • Review/adjust human review scope or frequency in response to system updates or newly emerging trends

H: Incorporating feedback

What criteria must be met?

5.36 Feedback mechanisms are in place to maintain or improve performance over time. This includes updating the proactive technology with diverse and up-to-date datasets to reflect evolving trends or emerging types of illegal content and/or content harmful to children

and/or integrating ongoing feedback from users and individuals working in content moderation into its development, while managing the risk of introducing additional bias.

Why is this important?

5.37 Proactive technology must adapt to remain effective and relevant in dynamic environments. Incorporating diverse and up-to-date datasets and structured feedback supports continual improvement, helps address emerging harms and/or changes to user behaviour and reduces the risk or performance degradation or bias over time.

What should providers do?

5.38 Collect and assess feedback from individuals working in content moderation, users or incident reviews. Ensure training or reference data is regularly updated to reflect new content types or user behaviours. Assess updates to ensure system performance is maintained or improved without introducing additional bias.

What type of questions might providers consider?

- What mechanisms are in place to capture relevant feedback?
- How often is the system updated or retrained?
- How are updates tested to confirm effectiveness and fairness?

Illustrative activities

Table 5.8: Three stages and illustrative activity for criteria H: 'Incorporating feedback'

Stage	Illustrative activity
Initial assessment	• Evaluate whether feedback from users and individuals working in content moderation is used in retraining/updating • Understand update frequency and governance
Testing and configuration	• Simulate updates with new data (such as new input types or flagged examples) • Check performance stability before/after updates • Establish pipelines for feedback integration
Deployment and ongoing monitoring	• Implement feedback processes for content flagged by individuals working in content moderation and users, and ensure that these processes are designed and monitored to mitigate bias • Plan update schedule to incorporate up to date data and address emerging trends • Monitor impact of updates to avoid introduction of bias