[Final Draft] Proactive Technology Guidance

Additional Safety Measures

Statement

Issue date: likely February 2027. This is a DRAFT published on 16 November to be issued when the associated Codes of Practice are issued.

Contents

Section

1. Overview 3
2. Introduction 4
3. The proactive technology criteria 7
4. The process of applying the proactive technology criteria 10
5. Understanding the proactive technology criteria 13

1. Overview

Navigating the guidance

1.1 The guidance consists of the following sections:

- Section 1: Overview
- Section 2: Introduction
- Section 3: The proactive technology criteria
- Section 4: The process of applying the proactive technology criteria
- Section 5: Understanding the proactive technology criteria

2. Introduction

Background to the guidance

Purpose of this guidance

2.1 This guidance aims to help service providers implement the measures that Ofcom recommends in the Code of Practice relating to the deployment of proactive technology (other than hash matching technology)[1] under the Online Safety Act 2023 (the Act).[2] The guidance provides additional detail and practical examples to help service providers configure and deploy such proactive technology so that it meets the proactive technology criteria. 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.

2.2 At this time, the guidance is intended to support all providers in scope of ICU PT 1, deployment of proactive technology to detect image-based CSAM. The guidance is relevant whether sourcing from a third-party provider, developing proactive technology in-house, or reviewing an existing deployment of a proactive technology. However, pending further work, we may revisit and revise this guidance. As a result, while the measure associated with this guidance is currently confirmed in relation to the purpose of detecting unknown image-based CSAM, we may extend the guidance to other kinds of harms over time.

2.3 Hash matching technologies are not in scope of the main body of this guidance. However, in our Guidance on appropriate proportion of human review for hash matching, we set out guidance in relation to the appropriate proportion of human review to be applied to content identified through hash matching technologies which have been specifically recommended in Codes of Practice for both user-to-user and search services.[3]

[1] Illegal Content Code of Practice for User-to-User Services: Proactive Technologies (Unknown Content)
[2] This guidance will be finalised following Ofcom issuing relevant amendments to our Codes. Prior to that, it is considered in draft.
[3] Our Guidance on appropriate proportion of human review for hash matching.

The proactive technology measure

Table 2.1: Measures and their application

Number	Measure	Application
ICU PT1	Providers should ensure that, where technically feasible, they use one or a combination of the specified technologies effectively to analyse relevant inputs for the purpose of detecting, or supporting detection of, unknown image-based CSAM content.	• Large user-to-user services that are medium or high risk of image-based CSAM • User-to-user services with more than 700,000 monthly UK users that are high risk of image-based CSAM • User-to-user services that are file-storage and file-sharing services of any size that are high risk of image-based CSAM.

ICU PT 1 – Deployment of proactive technology to detect image-based CSAM content

2.4 Service providers should use the specified technologies[4] as listed in our measure. In selecting, and throughout its lifecycle, providers should ensure the relevant technology meets the proactive technology criteria.[5] We have assessed that in principle the specified technologies are capable of being accurate, effective and lacking in bias when providers do this.

2.5 Where a provider already uses one (or a combination) of the specified technologies, it should ensure that this existing proactive technology meets the proactive technology criteria.

2.6 Where a provider finds that its existing proactive technology cannot meet the proactive technology criteria, we consider that providers should source from a third party, develop technology in-house, or a combination of these approaches.

2.7 Where a provider develops proactive technology in-house for the purpose of this measure, it should record the following:

- How it has had regard to the factors we outline for consideration in deciding to develop proactive technology in-house. These factors are set out in section 4 of the Proactive Technology Guidance.
- An 'implementation plan', including a timeline for development with key milestones, and a deadline for using one (or a combination) of the specified technologies in a way that meets proactive technology criteria.

[4] 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'.

[5] 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.

- How the timelines in its 'implementation plan' reflect a 'reasonable timeframe' as appropriate for their service.

2.8 As recommended in the measure, once a provider is using one or a combination of the specified technologies in a way that meets the criteria, it should maintain sufficient records to demonstrate the basis on which it concluded the relevant technology met the proactive technology criteria, including:

a) Records of any testing or evaluation undertaken;
b) The performance metrics used;
c) The results obtained;
d) configuration decisions;
e) identified limitations and mitigations; and
f) any ongoing monitoring and review activities relevant to its assessment.

2.9 At least every 12 months, or earlier where there is a change to any aspect of the service's design or operation which would amount to a significant change and therefore trigger a risk assessment under section 9(4) of the Act, the provider should review its selection of relevant technology.

Technical feasibility

2.10 The Act requires Ofcom to have regard to the principle that the measures described in Codes of Practice must be technically feasible. 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 of the service, its operation or both.

2.11 Where a provider has not deployed any proactive technology because it is not technically feasible to do so, we would expect it to record this decision and its reasoning. Where a service provider claims that it is technically infeasible for it to implement proactive technology for our measures, and we have concerns about this, we have the power to investigate. Should we find that it is technically feasible for the provider to implement the proactive technology, the relevant measure will apply to them. We do not consider that technical limitations will necessarily remain on an ongoing basis.

Interaction with data protection law

2.12 Nothing in this guidance, insofar as it relates to the processing of personal data, either supersedes or derogates from the requirements of applicable data protection legislation. Providers should continue to refer to the relevant guidance issued by the Information Commission's Office to ensure they are complying with their data protection obligations.

3. The proactive technology criteria

3.1 This guidance is intended to explain how a regulated service provider should consider the concepts of accuracy, effectiveness and lack of bias when considering the deployment of technology specified in Codes:

a) Accuracy: the technical correctness and reliability of 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.
b) Effectiveness: the utility and ability of 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.
c) Lack of bias: the extent to which 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.

The criteria

3.2 In having regard to accuracy, effectiveness and lack of bias, we have developed a set of criteria.

3.3 Where proactive technology can be deployed for the purpose of detecting target content 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.

3.4 The proactive technology criteria are as follows:

Table 3.1: Proactive technology criteria

Criteria	Description
High-quality data	The high-quality data criterion is fulfilled if the **relevant technology**[6] has been developed and tested using high-quality[7] datasets appropriate to and reflecting a broad range of inputs relevant to the harm it is intended to **detect** (as identified in the provider's **risk assessment**).

[6] 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.

[7] "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

Criteria	Description
Addressing biases	The addressing biases criterion is fulfilled if: a) potential biases have been identified and addressed; and b) risks are appropriately managed and addressed throughout the **relevant technology's** lifecycle.
Evaluating performance	The evaluating performance criterion is fulfilled if: a) the **relevant technology** has been evaluated using appropriate performance metrics; and b) configured so that its performance strikes an appropriate balance between **precision**[8] and **recall**[9].
Safeguards against misuse and exploitation	The safeguards against misuse and exploitation criterion is fulfilled if safeguards are in place to identify and appropriately manage security threats and risks of exploitation and misuse, including through the use of access restrictions[10] and system integrity protections.
Contextual testing and evaluation	The contextual testing and evaluation criterion is fulfilled if: a) the **relevant 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 provider's **risk assessment**); and b) the results indicate that it correctly **detects** the harm it is intended to **detect**.
Maintenance and ongoing monitoring	The maintenance and ongoing monitoring criterion is fulfilled if mechanisms are in place to monitor, maintain and quality assure the **relevant technology's** effectiveness over time, including processes for regular review, quality assurance and iterative adjustments to respond to emerging circumvention techniques, biases, or new content types.
Human review	The human review criterion is fulfilled if: a) policies and processes are in place for human review for both moderation and quality assurance purposes; and b) the provider abides by its policies and follows its processes.

(for example) unclear sourcing, evidence of poor or inconsistent labelling, or limited representation of relevant harms or input types.

[8] Precision: A measure of statistical accuracy, calculated as the proportion of detected content that a relevant technology has correctly identified as target content.

[9] Recall: A measure of statistical accuracy, calculated as the proportion of target content analysed by a relevant technology that the technology has detected.

[10] 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.

Criteria	Description
Incorporating feedback	The incorporating feedback criterion is fulfilled if feedback and quality assurance mechanisms are in place to maintain or improve performance of the relevant technology and associated systems and processes over time. This includes: a) identifying, correcting and responding to errors and performance issues; b) updating the relevant technology with diverse and up-to-date datasets to reflect evolving trends or emerging types of image-based CSAM content; and c) integrating ongoing feedback from users and individuals working in moderation[11] into its development, while managing the risk of introducing additional bias.

3.5 In practice, the proactive technology criteria should be understood as part of an ongoing process, in which activities are undertaken at different and overlapping stages that inform and strengthen one another over time. In particular, the 'Maintenance and Ongoing Monitoring' and 'Incorporating Feedback' criteria are closely interrelated. Insights generated through monitoring will inform system updates, refinements and retraining, and changes made through feedback mechanisms will in turn be evaluated through continued monitoring. Human review also plays an important role across both stages. However, it is set out as a separate section to ensure the guidance clearly addresses the different purposes and forms of human review that may be required across the end-to-end processes within which proactive technology is deployed. Although the criteria are presented separately to support clarity and practical application, this structure is not intended to suggest that they operate in isolation or in a linear sequence, and providers should consider how the criteria can be met collectively and continuously.

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

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

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

4. The process of applying the proactive technology criteria

4.1 This section outlines the process providers should follow to ensure relevant technology meets the proactive technology criteria.

4.2 For the purposes of this guidance, we have grouped activities in the lifecycle of proactive technology into three high-level stages:

a) initial assessment;
b) testing and configuration; and
c) 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 configuring, 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 in question. Whatever approach is taken providers should consider how each criterion can be met across the full lifecycle of the relevant 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 and an offence of possession of a prohibited image of a child. Providers should ensure they have taken appropriate legal advice on these offences (and relevant defences) and do not act unlawfully while implementing this measure, for example, in the course of compiling datasets on which proactive technology is trained. The Crown Prosecution Service and National Police Chiefs' Council have signed a Memorandum of Understanding to clarify the position of those involved in the management, operation or use of telecommunications services with regard such criminal offences and the defences available.[12]

4.6 Further details of how providers might apply the proactive technology criteria throughout the lifecycle of the relevant technology, including illustrative examples of practical steps, are provided in Section 5: Applying the proactive technology criteria.

4.7 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.

[12] Crown Prosecution Service, 2026. Memorandum of Understanding Between the Crown Prosecution Service (CPS) and the National Police Chiefs' Council (NPCC) concerning Section 46 Sexual Offences Act 2003 (updated May 2026) | The Crown Prosecution Service.

Initial assessment

Providers that choose to develop proactive technology in-house.

4.8 Providers which cannot meet the criteria through the configuration of their existing proactive technology may decide to develop proactive technology in-house to detect target content. Providers may take this approach provided the in-house technology is one of the specified technologies and meets the proactive technology criteria.

4.9 Providers which decide to develop proactive technology in-house do not need to conduct the initial assessment stage of the process but the proactive technology criteria should still be applied when testing, configuring, deploying and monitoring these solutions.

4.10 Where providers 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.

Providers that choose to explore third-party solutions

4.11 Providers exploring third-party solutions may 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.

4.12 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 from third parties at an early stage to form a reasonable judgement about whether the proactive technology is likely to be capable of being configured and deployed in a way that meets the criteria on their particular service. 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 deployment.

4.13 Where this initial assessment is carried out thoroughly and based on clear evidence, it should provide a strong indication that proactive technology can meet the criteria, reducing the risk of investing time and resources in a solution that ultimately does not meet the criteria.

Testing and configuration

4.14 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 relevant 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.

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

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

Deployment and ongoing monitoring

4.16 Once deployed, providers should ensure the relevant technology continues to meet the criteria over time through regular monitoring, quality assurance, 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.

4.17 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 relevant technology.
- Refine the relevant technology as or when necessary, so as to ensure that it meets the proactive technology criteria throughout its lifecycle.

5. Understanding the proactive technology criteria

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

- why it is important;
- what providers should do in order to configure and deploy proactive technology in a way that meets the criterion and the type of questions providers might consider; and
- illustrative examples of practical steps that may be taken to meet the criterion.

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 which of Ofcom's specified technologies to select and use in order to meet the proactive technology criteria in the context of the service in question. 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 our measure. However, the specific action a provider should take may differ depending on the context. For example:

- Providers **sourcing** proactive technology from a third party are likely to focus on gathering evidence from the third party provider and verifying it through testing and monitoring.
- Providers **developing** proactive technology will typically carry out all activities directly as part of their design and implementation process.
- Providers **configuring an existing deployment** of proactive technology may interpret the illustrative examples as prompts for considering 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 specified technologies with other technologies to ensure the relevant criteria are met.

5.5 In considering how to ensure relevant technology meets with the criteria, we expect providers to interpret and apply the criteria based on the context of their individual service, how the harm manifests on their service as set out in their risk assessment, and the characteristics of the proactive technology under consideration.

A: High-quality data

What criteria must be met?

5.6 The high-quality data criterion is fulfilled if the relevant 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 provider's risk assessment).

Why is this important?

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

5.8 If the datasets do not reflect the diversity of inputs the relevant 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.

5.9 We consider that a reasonably high level of expertise in data testing and training is required for providers to meet this criterion. Depending on the context, providers' own in-house expertise may be sufficient for this, and both third party review of data and in-house review of data can ensure quality assured datasets.

What should providers do?

5.10 Providers should assess whether the datasets used (for training, testing or operation) are sufficiently high-quality and appropriate to the harm and service context.

5.11 Providers should identify available mitigations which could be applied to address gaps or limitations.

5.12 Providers should use data from trusted flaggers to train their proactive technology where appropriate.

5.13 Where deploying technology that detects text, providers should consider the particular needs of their UK user base in relation to languages.

What type of questions might providers consider?

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

Illustrative activities

Table 5.1: Three stages and illustrative activity for criterion A: '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 relevant 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?

5.14 The addressing biases criterion is fulfilled if:

a) potential biases have been identified and addressed, and
b) risks are appropriately managed and addressed throughout the relevant technology's lifecycle.

Why is this important?

5.15 Providers should understand how bias has been or will be identified and addressed during initial assessment, testing and configuration, development, deployment and ongoing monitoring. Providers should consider how decisions made across the lifecycle of the relevant technology, such as in data selection, data labelling, rule design, or threshold setting (as relevant) may contribute to bias or unintended outcomes and identify steps to monitor and mitigate any disparities.

5.16 Bias introduced during any stage of the relevant technology's lifecycle can lead to discriminatory or unequal treatment of different user groups, types of content or other inputs if not addressed.

What should providers do?

5.17 Providers should understand how bias has been or will be identified and addressed during the lifecycle of the relevant technology.

5.18 Providers should consider how decisions made across the lifecycle of the relevant technology, such as in data selection, data labelling, rule design, or threshold setting (as relevant), may contribute to bias or unintended outcomes.

5.19 Providers should identify steps to monitor any disparities (see table 5.2). This should include monitoring any bias that the proactive technology has in relation to protected characteristics under the Equality Act 2010 and how this bias may affect the accuracy and fairness of outcomes.[13] Where the use of proactive technology may involve processing of special category data, often reflective of protected characteristics, providers should have regard to the guidance from the Information Commission on the safeguards required for processing of special category data in automated decision-making.[14]

5.20 Providers should identify steps to mitigate any disparities where appropriate, including the use of safeguards, improving training data representativeness, data augmentation and technology calibration, layering of technologies where there are higher errors.

What type of questions might providers consider?

- How has the relevant technology been tested across different user groups?
- How often is the relevant 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 relevant technology's decision-making process transparent?

Illustrative activities

Table 5.2: Three stages and illustrative activity for criterion 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 relevant 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

[13] Age, disability, gender reassignment, marriage and civil partnership, race, religion or belief, sex, sexual orientation.

[14] Information Commission's Office, 2026. Automated decision-making, including profiling.

C: Evaluating performance

What criteria must be met?

5.21 The evaluating performance criterion is fulfilled if:

a) the relevant technology has been evaluated using appropriate performance metrics, and
b) configured so that its performance strikes an appropriate balance between precision and recall.

Why is this important?

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

What should providers do?

5.23 Providers should understand how the performance of the tool is measured and calibrated and whether the metrics used are appropriate to the relevant technology under consideration.

5.24 Providers should assess whether the thresholds and settings are appropriate for the service context and risk the relevant technology seeks to address.

5.25 Providers should ensure the relevant technology is configured so that its performance strikes an appropriate balance between precision and recall. An appropriate balance of precision and recall is a balance set with a view to minimising the incidence and impact of moderation errors including both failures to identify relevant content and incorrect action against content. When configuring the technology so that it strikes an appropriate balance between precision and recall, providers should ensure that the following matters are taken into account:

- the service's risk of the relevant harm in scope of this measure, reflecting the risk assessment of the service and information reasonably available to the provider about the prevalence of target 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
- the importance of minimising the reporting of false positives to the National Crime Agency (NCA) or a foreign agency.

5.26 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 to privacy or legitimate political expression.

5.27 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 relevant technology is incapable of meeting the criteria. In such cases, providers in scope of the measure will need deploy another proactive technology model to comply with the measure.

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 relevant 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 criterion 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?

5.28 The safeguards against misuse and exploitation criterion is fulfilled if 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?

5.29 Proactive technology for content detection is susceptible to adversarial exploitation and misuse, potentially leading to compromised reliability.

5.30 Safeguards such as input validation[15], evasion detection[16] and configuration constraints[17] help ensure the relevant technology functions as intended and maintains performance. Safeguards such as access restrictions[18] and mechanisms to ensure system and data integrity are also required to address security risks and support reliable operation throughout the lifecycle of the relevant technology.

What should providers do?

5.31 Providers should identify what safeguards are in place to prevent exploitation or misuse, including adversarial resistance mechanisms and fail-safes.

5.32 Providers should understand the monitoring and alerting systems in place to identify unusual or unexpected behaviour.

5.33 Providers should evaluate incident response and escalation protocols.

What type of questions might providers consider?

- How does the relevant 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 relevant technology or its underlying data throughout its lifecycle.

---

[15] The technical processes and controls used to verify that data or content submitted to a content moderation system conforms to predefined requirements before being processed by the system. These controls seek to identify and appropriately handle invalid, malformed, manipulated, or potentially malicious inputs that could compromise the system's functionality, interfere with its ability to detect harmful content, or enable users to circumvent moderation safeguards

[16] The technical measures used to identify and respond to attempts to circumvent or undermine a content moderation system's ability to detect prohibited or harmful content. These measures seek to identify adversarial manipulation of inputs, suspicious patterns of activity, or other techniques intended to impact the effectiveness of the system

[17] Technical and administrative constraints governing how a content moderation system may be configured, modified, or operated. These constraints are designed in a threat-aware way which considers the specific harm manifestation and deployment context to ensure the effectiveness, security, or compliance of the content moderation system with established safety requirements

[18] "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.

Illustrative activities

Table 5.4: Three stages and illustrative activity for criterion 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 suppliers to patch vulnerabilities or adjust safeguard mechanisms as required

E: Contextual testing and evaluation

What criteria must be met?

5.34 The contextual testing and evaluation criterion is fulfilled if:

a) the relevant 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 provider's risk assessment), and
b) the results indicate that it correctly detects the harm it is intended to detect.

5.35 This evaluation includes testing for scalability, handling of different media types (where relevant), and assessing whether the relevant 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?

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

What should providers do?

5.37 Providers should conduct testing in conditions that reflect actual service use. This should include assessing how the relevant technology performs across relevant types of target content or other inputs and evaluating and mitigating any observed degradation in performance.

5.38 Proactive technologies are subject to general limitations which mean reliability can vary depending on the characteristics of the inputs being analysed. This means that the performance of a relevant technology may not be uniform across types of content, other inputs, or harm types and that the degree of confidence that can be placed in its outputs may vary accordingly.

5.39 The reliability of detection outputs is generally higher when:

- The inputs analysed clearly match a well-defined harm category, illegal content proxy or content harmful to children proxy;
- The inputs align with what the relevant technology has been developed and tested to detect or infer;
- The meaning, depiction, significance or risk indicated by the inputs is obvious, can be understood on its own, and further context or user history is not required;
- The relevant technology has demonstrated accurate past performance when analysing similar inputs.

5.40 The reliability of detection outputs is generally lower when:

- The relevant technology has not been developed or tested to analyse or detect that type of input;
- The meaning, depiction, significance or risk indicated by the inputs is ambiguous and additional contextual information is required to analyse and understand it;
- The relevant technology has a history of poor performance when analysing similar inputs.

5.41 Through testing, providers should evaluate the performance of the relevant technology including its accuracy and error profile.

5.42 Based on the evidence obtained from their evaluation of the relevant technology, providers should consider their level of assurance in the detection outcomes. The level of assurance reflects the degree of confidence that a provider can reasonably place in the system as a whole.

5.43 Level of assurance in the performance of the relevant technology will generally be high when the evaluation provides evidence that:

- The relevant technology performs accurately against relevant and representative ground-truth data aligned with its intended purpose.
- Performance of the relevant technology has been evaluated across relevant content types, contexts, and known edge cases, and no material gaps have been identified.
- System thresholds or configurations have been set on the basis of testing evidence to appropriately balance precision and recall, resulting in levels of false positives and false negatives that are proportionate to the risks associated with the relevant technology's intended use.
- Known limitations or failure modes have been identified and documented, and appropriate mitigations (such as human review) can be applied where required.
- Performance is stable and consistent, without significant unexplained variation across comparable inputs, testing instances, or operating conditions.

5.44 Level of assurance in the performance of relevant technology will generally be low when evaluation provides evidence that:

- Performance against relevant ground-truth data is inconsistent, insufficiently evidenced, or unreliable, particularly in scenarios where higher reliability would reasonably be expected.
- Performance varies materially across content types, contexts, or edge cases in ways that are not well understood or cannot be adequately explained.
- Available testing evidence is not sufficient, or it has not otherwise been possible to establish a balance between precision and recall that would result in levels of false positives and/or false negatives proportionate to the risks associated with the relevant technology's intended use
- Important limitations or failure modes remain unidentified, poorly characterised, or cannot be reasonably mitigated within the proposed deployment model.
- The system exhibits instability or unpredictable behaviour, including unexplained fluctuations in performance under comparable conditions.

5.45 Where only a low level of assurance can be established, and identified limitations cannot be effectively mitigated, this may indicate that the relevant technology is not capable of meeting the proactive technology criteria on the service.

5.46 Level of assurance in the performance of the relevant technology should be treated as being materially reduced where any of the following are present:

a) Significant inconsistencies between layered proactive technology tools or signals from other detection tools or metadata analysis, when analysing the same or similar inputs during testing;
b) Significant inconsistencies between relevant technology outputs and outcomes of quality assurance such as human review;
c) Conflicting signals across rule-based checks and contextual or behavioural analysis, where predefined rules flag inputs as being suspected target content based on indicators (such as keywords or history of user behaviour), but other tools assessing the meaning or context of the content or user behaviour indicate that the inputs do not relate to activity that would render it target content;
d) Novel input formats, emerging trends or indications of adversarial behaviour which the relevant technology has not been trained to detect.

What type of questions might providers consider?

- Can the relevant 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 criterion 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 relevant 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?

5.47 The maintenance and ongoing monitoring criterion is fulfilled if mechanisms are in place to monitor, maintain and quality assure the relevant technology's effectiveness over time, including processes for regular review, quality assurance and iterative adjustments to respond to emerging circumvention techniques, biases, or new content types.

Why is this important?

5.48 Ongoing monitoring and adjustment to the relevant technology are essential to maintain the effectiveness of the system over time and to prevent degradation of performance as content evolves, new manifestations of harm emerges and/or user behaviour changes.

What should providers do?

5.49 Providers should establish monitoring and quality assurance processes to track performance of the relevant technology and identify and respond to emerging risks. Monitoring and quality assurance should be carried out regularly and on an ongoing basis to ensure that findings remain relevant and actionable.

5.50 Providers should use quality assurance to:

a) identify the prevalence, causes and patterns of false positives and false negatives;
b) assess whether the relevant technology is appropriately calibrated, as evidenced by observed error rates, performance variation, and outcomes from quality assurance;
c) identify inconsistencies in the performance of the relevant technology across harm categories, content types or user groups;

d) identify any novel content formats, emerging trends or indications of adversarial behaviour which the relevant technology has not been trained to detect;
e) assess whether detection errors indicate limitations in training datasets, configuration of the relevant technology and related systems or human review processes.

5.51 Quality assurance should involve sampling of:

a) detected content at different points in the content moderation pipeline, including:
    i) cases where a moderation action was taken (including both automated and human decisions); and
    ii) cases where no moderation action was taken (including both automated and human decisions);
b) content which may be a false negative, identified through user reporting, complaints, community moderating, stratified sampling or other means; and
c) high-impact and high-uncertainty cases (including edge cases and emerging content types).

5.52 Quality assurance should take account of factors that may affect detection accuracy, such as:

a) low-quality inputs (for example, blurred images, distorted audio or truncated text);
b) highly contextual content (including slang, irony or coded language);
c) different languages;
d) emerging content formats; or
e) adversarial behaviour which the relevant technology has not been trained to detect.

5.53 Providers should be aware that a randomly selected sample of the content on the service may only identify a small proportion of false positives and false negatives. Providers should consider whether additional and/or targeted sampling is needed to ensure meaningful quality assurance of the relevant technology.

5.54 Providers should ensure there is human review of an appropriate proportion of the sample for quality assurance, particularly in complex or nuanced cases. Further guidance on conducting human review for the purposes of quality assurance is set out in the section entitled ‘G: Human Review’.

5.55 Providers should incorporate secondary validation as part of their quality assurance, particularly in respect of outputs of the sample which cannot be validated by humans. This may include comparing outputs across different automated systems (for example, between different models, rule-based systems or third-party tools) to identify inconsistencies or anomalies. It could also include retesting the relevant technology using alternative datasets or scenarios (such as previously unseen inputs, new languages, edge or adversarial cases) to assess whether outputs of the relevant technology remain consistent and reliable.

5.56 Providers should update the relevant technology as necessary to respond to identified changes in performance. Further guidance on incorporating feedback is set out in the section entitled ‘H: Incorporating Feedback’.

What type of questions might providers consider?

- How is ongoing performance monitored and assessed?

- Has there been a review and/or secondary validation of a diverse sample for quality assurance?
- What triggers a review or an update of the relevant technology?
- How is, or can, the relevant technology be adapted to address newly identified risks?

Illustrative examples

Table 5.6: Three stages and illustrative activity for criterion 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 relevant technology based on observed drift or circumvention • Log updates and regularly review effectiveness

G: Human review

What criteria must be met?

5.57 The human review criterion is fulfilled if:

a) policies and processes are in place for human review for both moderation and quality assurance purposes, and
b) the provider abides by its policies and follows its processes.

Why is this important?

5.58 Human oversight is an essential and well-recognised safeguard ensuring that overall outputs of content moderation processes which use automated technology 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 technological outputs are explainable (meaning, it is possible to understand how and why they were generated), human reviewers are better able to assess and make informed decisions.

What should providers do?

5.59 Providers should ensure human review is incorporated into the development and testing process to assess early outputs and inform adjustments to the relevant technology.

5.60 Providers should also ensure an appropriate proportion of outputs are subject to human review during deployment.

5.61 There are two purposes for human review:

a) Moderation – providers may need to use human review to determine whether the content is target content, and
b) Quality assurance – using human moderators for quality assurance and monitoring of the relevant technology.

5.62 We provide further guidance on both forms of human review in this section.

5.63 As set out in our Codes, the following factors should be taken into account by providers when determining the appropriate proportion of outputs and undetected content (where applicable), subject to human review:

- the level of assurance the provider has in the detection outcomes produced by the technology and any associated systems and processes. This assurance can be determined by ongoing monitoring, evaluation and quality assurance (including human quality assurance) of the performance of the technology;
- the potential severity of the harm to those depicted in or encountering target content; and
- for content moderation purposes, the overall impact of an incorrect decision that the content is target content, on the user who generated, uploaded or shared the content, having particular regard to the importance of minimising the reporting of false positives to the NCA or a foreign agency.

5.64 We recognise that some of these factors may be in tension with one another. The appropriate balance between these factors, which will most effectively protect users on a service, may vary between services and harms. However, it should be determined having regard to the guidance below.

5.65 Providers should allocate human review resources on the basis of documented performance evidence and should review and adjust this allocation regularly to reflect observed accuracy, error patterns and risks of harm.

Human review of detected content for the purposes of moderation

5.66 Human review of detected content during deployment is an important safeguard for accuracy in a decision-making process that would otherwise have significant effects if the decision were wrong. It can also be used, in a proportionate way, to ensure there is fairness in a decision-making process that could otherwise rely solely upon proactive technology and supports compliance with providers' obligations under data protection legislation in respect of automated decision-making.

5.67 Providers should ensure human moderators review and assess an appropriate proportion of outputs of the proactive technology during deployment for moderation purposes. To determine what proportion of detected content is appropriate for human review before a relevant content moderation action is made, providers should take into account:

a) their level of assurance in the detection outcomes produced by the relevant technology and any associated systems and processes. For example, at lower levels of assurance more human review is likely to be appropriate;

b) the potential severity of the harm to those depicted in or encountering target content. For example, where human review may lead to a delay in content take down and the provider has a high level of assurance in the detection outcomes, it may be appropriate to have a lower proportion of human review;
c) the overall impact of an incorrect decision that the content is target content, on the user who generated, shared or uploaded the detected content, for example:

i) restrictions on a user's access to the service, or their ability to use the service;
ii) an adverse impact on the user's financial remuneration payable in connection with, or generated by, the detected content (such as advertising revenue, subscriptions, commission or other monetisation arrangements), where known, and
iii) having particular regard to the importance of minimising the reporting of false positives to the NCA or a foreign agency.

5.68 In considering detected content which may have a high impact on the user, the provider should also consider the extent to which and how promptly the relevant content moderation actions can be reversed.

Human review for quality assurance purposes

5.69 Human review of detected content for quality assurance may involve reassessing a sample of detected content and content moderation decisions relating to such content (including both automated and human decisions). The purpose of this type of human review is not to determine individual outcomes of moderation, complaints or appeals in real time, but to evaluate how the relevant technology and related systems perform in practice, identify any systemic issues and correct any identified errors.

5.70 More human review for the purposes of quality assurance is likely to be needed where the level of assurance the provider has in the detection outcomes produced by the technology and any associated systems and processes – including other layers of proactive technology and human moderators – is lower.

5.71 Providers should also increase human review as part of quality assurance where observed outcomes are inconsistent with expectations.

5.72 A proper approach to the configuration and ongoing quality assurance of relevant technology also requires human review of undetected content. This may include content which was not detected by the relevant technology but has been identified through user reporting, complaints, community moderation, stratified sampling or other means. Where such content is subsequently assessed as target content that the relevant technology was deployed to detect under its current configuration, this constitutes a false negative.

5.73 When deciding the appropriate proportion of detected content for human review as part of quality assurance, providers should have regard to the potential impact on users or depicted persons arising from a failure to detect target content, including the risk that target content remains available to be encountered by UK users.

5.74 When deciding the appropriate proportion of detected content for human review as part of quality assurance, providers should have regard to the potential impact on users or depicted persons arising from a failure to detect target content, including the risk that target content remains available to be encountered by UK users, or that content harmful to children remains available to be encountered by UK child users.

What type of questions might providers consider?

- What role does human review play in testing and refinement of the relevant technology prior to deployment?
- What proportion of outputs will be reviewed during deployment and how will this be determined?
- What role does human review play in the quality assurance of the relevant technology throughout its lifecycle.

Illustrative activities

Table 5.7: Three stages and illustrative activity for criterion 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?

5.75 The incorporating feedback criterion is fulfilled if feedback and quality assurance mechanisms are in place to maintain or improve performance of the relevant technology and associated systems and processes over time.

5.76 This includes:

a) identifying, correcting and responding to errors and performance issues,
b) updating the relevant technology with diverse and up-to-date datasets to reflect evolving trends or emerging types of image-based CSAM content, and
c) integrating ongoing feedback from users and individuals working in moderation into its development, while managing the risk of introducing additional bias.

Why is this important?

5.77 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 of performance degradation or bias over time.

What should providers do?

5.78 Providers should collect and assess feedback from quality assurance, individuals working in content moderation, users or incident reviews. Providers should have mechanisms in place to obtain feedback to understand whether there are any risks or emerging trends relating to the accuracy, effectiveness and/or lack of bias of the relevant technology, including by:

a) monitoring outcomes and analysing patterns in:

i) quality assurance;
ii) outcomes of human review of detected content;
iii) outcomes of complaints and appeals relating to detected content and identified false negatives;
iv) overturned or corrected moderation decisions;
v) overturn rates;

b) maintaining and tracking overturn rates and service-specific targets for:

i) maximum acceptable false positive/false negative prevalence in relevant content moderation actions and user sanctions;
ii) the time taken to reclassify the content and restore the content (where applicable).

5.79 Providers may identify false positives and false negatives during the course of obtaining feedback such as through quality assurance, content moderation or human review.

5.80 Where a false positive has been identified, the provider should:

a) recategorise the content correctly;
b) so far as appropriate and possible, restore the content on the service so that it is available to:

i) all UK users (if the content is lawful and not harmful to children);
ii) adult UK users (if the content is lawful but is primary priority content harmful to children);

c) so far as appropriate and possible, reverse any user sanctions that have been made as a result of the incorrect classification;
d) record the false positive and update any relevant metrics accordingly, such as the overturn metrics.

5.81 Where a false positive has been identified, the provider should:

a) recategorise the content correctly;
b) record the false positive and update any relevant metrics accordingly, such as the overturn metrics.

5.82 Where a false negative has been identified, the provider should:

a) recategorise the content correctly;
b) remove the content as follows:

i) if the content is illegal content, swiftly take it down
ii) if the content is lawful but is content harmful to children, make it unavailable to UK child users;

c) apply relevant user sanctions under the provider's user sanctions policy;

d) record the false negative and update any relevant metrics accordingly, such as the overturn metrics.

5.83 Providers should consider whether feedback indicates a disproportionate level and/or pattern of false positives and false negatives. For example, whether false negatives are concentrated in particular harm categories or content types, or where there is an emerging trend of false positives, particularly if it affects forms of political expression. Your feedback loop may also indicate the need to consider content in particular languages, see the section entitled 'High-quality data'.

5.84 Providers should ensure training or reference data is regularly updated to reflect new content types or user behaviours. Providers should assess updates to ensure system performance is maintained or improved and to identify any relevant effects on bias.

5.85 Providers should incorporate feedback into their relevant technology and related systems promptly and on an ongoing basis. This enables identified errors or limitations to inform updates, recalibration, reconfiguration or retraining of the relevant technology (including through engagement with third-party providers of proactive technology, where relevant).

5.86 Providers should also incorporate feedback as appropriate into any relevant practices, processes, guidance and criteria for the wider end-to-end process within which relevant technology is deployed.

5.87 When incorporating feedback, providers should prioritise updates based on the impact and frequency of observed errors and limitations. Where appropriate, providers should consider staged, incremental or otherwise controlled deployment of material changes, so that unintended performance effects can be identified and attributed before wider deployment.

5.88 Following the incorporation of feedback, providers should re-evaluate the level of assurance and configuration settings of the relevant technology to ensure that, following any updates or changes, the relevant technology is sufficiently accurate, effective and lacking in bias.

5.89 Providers should keep a record of any changes made to the relevant technology as a result of feedback, together with the reasons for those changes.

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 criterion 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

Stage	Illustrative activity
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