Accreditation for terrorism/CSEA content detection technology

Annex 10: [DRAFT] Minimum Standards of Accuracy in the Detection of Terrorism and CSEA Content

Statement

Published: 8th May 2026

Contents

Annex

A10. [DRAFT] Minimum Standards of Accuracy..... 3

A10. [DRAFT] Minimum Standards of Accuracy in the Detection of Terrorism and CSEA Content

Introduction

Legislative framework

A10.1 The Online Safety Act 2023 ('the Act') protects adults and children online. It puts a range of new duties on the providers of certain internet services,[1] such as social media and search services, including to protect their users from illegal content and content harmful to children. The Office of Communications ('Ofcom') is the independent regulator for online safety.

A10.2 Terrorism content[2] and child sexual exploitation and abuse ('CSEA') content[3] are both categories of illegal content. Chapter 5 of Part 7 of the Act gives Ofcom the power to issue a notice to the provider[4] of a regulated user-to-user service[5] or regulated search service[6] where it considers it necessary and proportionate to deal with terrorism content or CSEA content, or both (a 'Technology Notice').

A10.3 A Technology Notice could require a provider to:

a) use accredited technology to identify and/or prevent individuals from encountering[7] terrorism content communicated publicly[8]; and/or
b) use accredited technology to identify and/or prevent individuals from encountering CSEA content communicated publicly or privately or, alternatively, use best endeavours to develop or source technology that meets minimum standards of accuracy to deal with such content.[9]

[1] "Internet service" is defined in section 228 of the Act.
[2] "Terrorism content" is defined in section 59(8) of the Act as content that amounts to an offence specified in Schedule 5. "Content" is defined in section 236(1) of the Act. Section 59(3) sets out when content amounts to an offence. Section 192 sets out how, where they are required to do so, providers of services should make judgements as to whether content is illegal content.
[3] "CSEA content" is defined in section 59(9) of the Act as content that amounts to an offence specified in Schedule 6.
[4] "Provider" is defined in section 226 of the Act.
[5] "User-to-user service" and "regulated user-to-user service" are defined in sections 3 and 4 of the Act.
[6] "Search service" and "regulated search service" are defined in sections 3 and 4 of the Act.
[7] "Encounter" is defined in section 236(1) of the Act.
[8] See section 232 of the Act which specifies the factors that Ofcom must consider when deciding whether content is communicated "publicly" or "privately".
[9] For completeness, a user-to-user service may be required to use accredited technology to identify and swiftly take down, or prevent individuals from encountering, terrorism content or CSEA content. Search services may be required to use accredited technology to identify search content of the service that is terrorism content or CSEA content and swiftly take measures to secure that, so far as possible, search content no longer includes such content identified by the technology. "Search content" is defined in section 57(2) of the Act.

A10.4 'Accredited technology' is defined in the Act as technology that has been accredited, by Ofcom or another person appointed by Ofcom, as meeting minimum standards of accuracy in the detection of terrorism or CSEA content (as the case may be).[10] Those minimum standards must be such standards as are for the time being approved and published by the Secretary of State, following advice from Ofcom.[11]

A10.5 Following advice from Ofcom pursuant to section 125(13) of the Act,[12] the minimum standards of accuracy set out in this document ('the Minimum Standards') have been approved and published by the Secretary of State for the Department of Science, Innovation and Technology. They set out minimum standards of accuracy in the detection of terrorism and CSEA content (as the case may be) for the purposes of Chapter 5 of Part 7 of the Act.

A10.6 The Minimum Standards come into force on [DATE].

A10.7 Ofcom, or another person appointed by Ofcom, is ultimately responsible for determining whether a technology meets the Minimum Standards.

A10.8 Accreditation against the Minimum Standards does not guarantee that the requirements of other relevant legislation have been met by a technology. It also does not mean that Ofcom will consider it necessary and proportionate to require the use of that technology in a Technology Notice.[13]

A10.9 Neither does accreditation against the Minimum Standards signify that Ofcom (or any other regulator) has approved a technology or endorsed its use. Accreditation of a technology against the Minimum Standards is only intended to determine whether a technology could be considered for requirement through a Technology Notice.

Overview of the Minimum Standards

A10.10 The Minimum Standards are set out in Section 1. They are based on an audit-based assessment, consisting of four main Principles each of which is underpinned by outcomes-based Objectives.

A10.11 The four Principles are:

- Technical Performance
- Fairness
- Robustness
- Maintainability

A10.12 The Objectives, and the Principles to which they relate, are set out in Section 2. Additional information is included under 'Additional notes' in some cases.

[10] Section 125(12) of the Act.
[11] Section 125(13) of the Act.
[12] See Ofcom's [Research, evidence, and advice to the DSIT Secretary of State on how to set minimum standards of accuracy].
[13] Ofcom published guidance for providers of regulated user-to-user and regulated search services about how they propose to exercise their Technology Notice functions in May 2026. This guidance sets out the process Ofcom would typically follow when deciding whether it is necessary and proportionate to issue such a notice and some detail on the matters to which Ofcom would expect to have regard when making this decision. See Ofcom's [Guidance on the exercise of Ofcom's functions under Chapter 5 of Part 7 of the Online Safety Act 2023].

A10.13 Any technology developer seeking accreditation of their technology will be asked to submit evidence demonstrating how they meet each Objective and the overarching Principle. This evidence is gathered through a set of questions that are designed to support the assessment of each Objective (as applicable) and its corresponding Principle. These questions are not part of the Minimum Standards and will be set by Ofcom.

A10.14 The evidence submitted in response to these questions will be assessed and graded by Ofcom, or another person appointed by Ofcom, based on the following scale:

a) 5 points where evidence submitted is robust and comprehensive;
b) 1 point where evidence is submitted but it has notable limitations; or
c) Zero points where the evidence required for 1 or 5 points is not provided or if evidence provided appears to contain misleading or inaccurate information.

A10.15 These grades will be used to determine whether the technology has met the Minimum Standards in accordance with the scoring framework set out in Section 3.

A10.16 In order to be accredited, each of the Minimum Standards in Section 1 must be met.

Interpretation of terms used

A10.17 The following terms are used in the Minimum Standards and in associated documentation. In these Minimum Standards, unless the contrary intention is expressed, terms defined in the Act have the same meaning as in the Act.

- '**Adversarial attack**' involves manipulating input data to deceive the technology in making incorrect outputs, predictions, or classifications (for example, evasion attacks, injection attacks, and input perturbations).
- '**Deployment**' refers to an operational technology being put into use on a particular internet service. References to deploy shall be construed accordingly.
- '**Explainability**' involves post-hoc methods used to analyse how a machine learning-based technology produces its outputs. These explanations are generated after training and can take various forms, such as visualisations, feature importance plots, or textual explanations.
- '**Fairness**' refers to the ability of a technology to avoid unfair bias across different groups of people.
- **'False positive'** is incorrectly classifying a negative sample as positive.
- '**False positive rate**' refers to the proportion of negative cases incorrectly predicted by a technology as positive cases, calculated by dividing the number of false positives by the total number of actual negative cases (false positives and true negatives).
- '**Interpretability**' refers to when the behaviours and decisions made by a technology can be easily understood by humans. A technology is interpretable when its structure or operation is inherently understandable, or sufficient documentation makes its structure or operation clear.
- '**Maintainability**' refers to the ability of a technology to be modified, repaired, or updated to ensure its continued accuracy and performance over time, including in response to new threats.
- The **'Overall Score'** refers to the score calculated in accordance with paragraph A10.40 of the scoring framework. The method for calculating the Overall Score is based upon

applying a weighting of 30% to the scores for each of the Technical Performance, Fairness and Robustness Principles and 10% for the Maintainability Principle.

- '**Robustness**' refers to the ability of a technology to perform reliably and maintain functionality under various conditions, including unexpected or challenging scenarios.
- '**Technical Performance**' refers to the testing and reporting of a technology's ability to perform against specified metrics and technical requirements across comprehensive and representative datasets.
- **'True negative'** is correctly classifying a negative sample as negative.

Section 1: The Minimum Standards

A10.18 A technology must achieve a score of:

a) at least 80 out of 100 for Objective 1.1: Performance Metrics; and
b) at least 40 out of 100 for each of the other Objectives.

A10.19 A technology must achieve a score of at least 60 out of 100 for each Principle.

A10.20 A technology must achieve an Overall Score of at least 60 out of 100.

Section 2: The Principles and Objectives

Technical Performance

A10.21 **Objective 1.1: Performance Metrics.** The technology's ability to identify terrorism or CSEA content (as the case may be) has been comprehensively evaluated against the false positive rate and other appropriate performance metrics. Corresponding evaluation results are provided and demonstrate that the technology is able to detect terrorism or CSEA content (as the case may be), and those results have been used to determine that the technology is suitable for deployment in the environment(s) for which it has been designed.

A10.22 **Objective 1.2: Dataset Quality.** The datasets used in development, including where applicable the training and testing of the technology's performance, are sufficiently comprehensive, representative of the harm being detected and, where relevant, sufficiently diverse to test the technology's ability to generalise to data not seen during development.

A10.23 **Objective 1.3: Reproducible Performance.** The technology's performance is sufficiently consistent and reproducible across the environment(s) for which it has been designed.

A10.24 **Objective 1.4: *Secondary Validation.** The technology's outputs, where possible, have been evaluated during performance testing against expert human judgement, particularly in complex or nuanced cases. Where outputs cannot be validated by humans, other secondary validation measures have been undertaken.

Additional notes

a) Objective 1.4: Secondary Validation is marked with an asterisk (*). This is a specific Objective for which it may be challenging for some technology developers to provide evidence. Specifically, where technologies have been developed without access to input/output and/or training data. In these cases, the technology developer seeking

accreditation should provide evidence against this Objective (and underling questions) to the extent possible.[14]

b) Where a technology developer nevertheless remains unable to provide any evidence against the Objective (or any of the underlying questions), it should confirm to Ofcom's satisfaction (or another person appointed by Ofcom) that its technology was developed without access to input/output and/or training data and explain why it has been unable to provide any relevant evidence (and the steps taken to assure itself of this).

c) If Ofcom (or another person appointed by Ofcom) is satisfied that a technology developer has a valid reason for being unable to provide evidence against any of the questions and/or the Objective, the relevant question(s) and/or Objective will be excluded for the purposes of calculating the scores used to determine whether the technology meets the Minimum Standards in accordance with the scoring framework set out in Section 3.

Fairness

A10.25 **Objective 2.1: *Bias Identification and Mitigation.** Comprehensive policies, procedures, metrics, and analyses have been implemented to identity potential biases in the technology throughout planning and development. In addition, robust bias mitigation strategies have been implemented and their success has been measured over time, including checks for demographic fairness and audits on any automated decision making.

A10.26 **Objective 2.2: *Data Processing.** The data processing used for any relevant training or testing datasets is robust, with documented, standardised criteria used to process data. Measures have also been taken to ensure consistency and minimise bias and errors during the processing.

A10.27 **Objective 2.3: Interpretability and Explainability.** The rationale behind algorithmic decisions made by the technology can be sufficiently understood by Ofcom and companies that are likely to deploy the technology.

Additional notes

a) Objective 2.1: Bias Identification and Mitigation and Objective 2.2: Data Processing are marked with an asterisk (*). These are specific Objectives for which it may be challenging for some technology developers to provide evidence. Specifically, where technologies have been developed without access to input/output and/or training data. In these cases, the technology developer seeking accreditation should provide evidence against those Objectives (and underling questions) to the extent possible.[15]

b) Where a technology developer nevertheless remains unable to provide any evidence against one or both Objectives (or any of the underlying questions), it should confirm to Ofcom's satisfaction (or another person appointed by Ofcom) that its technology was developed without access to input/output and/or training data and explain why it has been unable to provide any relevant evidence (and the steps taken to assure itself of this).

c) If Ofcom (or another person appointed by Ofcom) is satisfied that a technology developer has a valid reason for being unable to provide evidence against any of the

[14] Where the technology developer does not have direct access to relevant information required to respond (e.g., the developer cannot view the datasets against which the technology was developed or tested), it should explore alternative ways of providing relevant evidence.

[15] See footnote 14.

questions and/or Objectives, the relevant question(s) and/or Objective(s) will be excluded for the purposes of calculating the scores used to determine whether the technology meets the Minimum Standards in accordance with the scoring framework set out in Section 3.

Robustness

A10.28 **Objective 3.1: Development in a Secure Environment.** The technology has been developed with sufficient cybersecurity, privacy, and data protection measures in place, particularly for ensuring the integrity of the algorithm and protection of sensitive data. Documentation of how secure design principles have been followed during software development is provided.

A10.29 **Objective 3.2: Consistent Performance Over Time.** The technology maintains expected operation and performance over time, demonstrating its reliability and stability in tests representative of the environments for which it has been designed (and, where relevant, when deployed). Any degradation over time is monitored and reported on.

A10.30 **Objective 3.3: Robust Incident Handling and Recovery.** The technology includes robust incident handling and recovery mechanisms, enabling the management of system failures or unexpected situations.

A10.31 **Objective 3.4: Reliable Operation Across Relevant Services, Devices, and System Demands.** The technology operates reliably in tests representative of the services and devices it was designed to operate on (and, where relevant, when deployed), and varying system capacity demands.

A10.32 **Objective 3.5: Detection and Mitigation of Threats.** Sufficient safeguards and processes are in place to detect and mitigate both intentional and unintentional threats, which may include input manipulation and contextual misunderstandings. The technology can effectively respond to a wide range of adversarial attacks and circumventions of intended use while maintaining its integrity and accuracy.

Maintainability

A10.33 **Objective 4.1: System Risk and Update Management.** Comprehensive procedures and policies are in place for proactive identification and management of system-level risks, with a view to ensuring that the performance of the technology is maintained both over time, and across subsequent updates.

A10.34 **Objective 4.2: Effective Quality Assurance (QA) Plans and Periodic Monitoring.** Effective Quality Assurance (QA) policies are in place to address organisational risk and procedural oversight, and periodic monitoring procedures are conducted with a view to ensuring the maintenance or improvement of the technology's performance. The processes for development and maintenance of the technology over time have been documented.

A10.35 **Objective 4.3: Data Lifecycle and Retention Governance.** Comprehensive procedures and policies are in place to govern the retention, archiving, disposal, and general management of data relating to the operation of the technology (including both data used to develop the technology as well as data about the technology's development) with a view to ensuring operational consistency across the technology's lifecycle.

A10.36 **Objective 4.4: Stakeholder Feedback Incorporation.** The technology provider has processes in place to incorporate customer feedback into the ongoing monitoring and evaluation of the technology's performance.

Section 3: The scoring framework

A10.37 This Section sets out how to calculate the scores used to determine whether a technology meets the Minimum Standards.

How to calculate the score for each Objective

A10.38 A technology's score for each Objective is calculated as follows:

Step 1 – aggregate the grades for each question

Aggregate the grades for each question that relates to the Objective to calculate the total score achieved for the Objective.

Step 2 – calculate the maximum potential score for the Objective

Calculate the maximum potential score for the Objective by multiplying the number of questions which relate to the Objective by 5.

Step 3 – divide the total score achieved by the maximum potential score for the Objective

Divide the figure calculated under Step 1 by the figure calculated under Step 2.

Step 4 – convert to a score out of 100

Multiply the figure calculated under Step 3 by 100 to convert to a score out of 100.

The technology's score for the Objective is the figure as calculated under Step 4.

How to determine the score for each Principle

A10.39 A technology's score for each Principle is calculated as follows.

Step 1 – aggregate the grades for each question

Aggregate the grades for each question that relates to the Principle to calculate the total score achieved for the Principle.

Step 2 – calculate the maximum potential score for the Principle

Calculate the maximum potential score for the Principle by multiplying the number of questions which relate to the Principle by 5.

Step 3 – divide the total score achieved by the maximum potential score for the Principle

Divide the figure calculated under Step 1 by the figure calculated under Step 2.

Step 4 – convert to a score out of 100

Multiply the figure calculated under Step 3 by 100 to convert to a score out of 100.

The technology's score for the Principle is the figure as calculated under Step 4.

How to calculate the Overall Score

A10.40 The Overall Score is calculated as follows –

(T x 0.3) + (F x 0.3) + (R x 0.3) + (M x 0.1)

where –

T is the score for the Technical Performance Principle,

F is the score for the Fairness Principle,

R is the score for the Robustness Principle, and

M is the score for the Maintainability Principle.