Technology Notices to deal with terrorism content and/or CSEA content

Consultation on policy proposals for minimum standards of accuracy for accredited technologies, and guidance to providers

Consultation

Published: 16 December 2024
Closing date for responses: 10 March 2025

Contents

Section

1. Overview

1.1 Ofcom is the United Kingdom's (UK) communications regulator, overseeing sectors including telecommunications, post, broadcast TV, radio, and online services. We were appointed the online safety regulator under the Online Safety Act 2023 ('the Act') in October 2023.

1.2 Under the Act, the providers of regulated user-to-user and regulated search services ('Part 3 services') have a range of new duties.[1] They must assess the risk of harm arising from illegal content or activity on their regulated services and take or use proportionate measures to effectively manage and mitigate those risks. They must also take or use proportionate measures to prevent individuals from encountering 'priority' illegal content by means of their regulated services.

1.3 Ofcom has published its Illegal Harms Codes of Practice which set out recommended measures that service providers can take to comply with their illegal content safety duties. Ofcom can take enforcement action where providers breach their illegal content safety duties, which may include where they do not take recommended measures or appropriate alternative measures to comply. Ofcom has also published its Online Safety Enforcement Guidance which explains when and how Ofcom will consider taking enforcement action.

1.4 Ofcom also has additional powers to tackle two categories of illegal content: terrorism and child sexual abuse and exploitation (CSEA). Under section 121 of the Act, we have the power to issue a notice to the provider of a particular Part 3 service where we consider it necessary and proportionate to deal with terrorism or CSEA content (or both). We refer to such a notice as either a Technology Notice or a Notice throughout this publication. This Notice could require a Part 3 service provider to:

- use technology that has been accredited ('accredited technology'), by Ofcom or another person appointed by Ofcom to identify and/or prevent individuals from encountering terrorism content communicated publicly;[2] and/or
- use accredited technology to identify and/or prevent individuals from encountering CSEA content communicated publicly or privately. Alternatively, the Notice could require a regulated service provider to use best endeavours to develop or source technology that meets the minimum standards or accuracy to deal with such CSEA content, rather than requiring use of an already accredited technology.

[1] Regulated user-to-user and regulated search services are defined in the Act as 'Part 3 Services' because Part 3 of the Act imposes duties on providers of these services. We have adopted this definition throughout this consultation.

[2] The wording of the Act distinguishes between how Technology Notices may apply to regulated user-to-user services and regulated search services. For regulated user-to-user services, a notice may require a provider to use accredited technology to identify and swiftly take down or prevent individuals from encountering terrorism and/or CSEA content. For regulated search services, a notice may require a service to: use accredited technology to identify search content of the service that has terrorism and/or CSEA content; and to swiftly take measures designed to make sure that, as far as possible, search content of the service no longer includes terrorism and/or CSEA content identified by the technology. See section 121 (2) and (3) for more information on the provisions for regulated user-to-user and regulated search services respectively.

1.5 Before Ofcom can consider issuing a Technology Notice, some important steps need to have been taken:

- Ofcom must advise the Secretary of State about minimum standards of accuracy in the detection of terrorism content and CSEA content before the Secretary of State can approve and publish them;
- Before it can issue a Notice requiring the use of a specific technology, Ofcom or a nominated third party must have accredited that technology against the minimum standards of accuracy; and
- Ofcom must produce guidance for Part 3 service providers about how it proposes to exercise its Technology Notice functions.

What this document covers

1.6 Ofcom is consulting on its policy proposals to the Secretary of State on setting minimum standards of accuracy. We are taking this step to make sure we have the best possible evidence to inform our advice and so are seeking views from a range of stakeholders on our proposals. This includes providers of regulated services, technology developers, researchers, academics, and organisations focused on making the UK a safer place to be online. This is particularly important as the evaluation of safety technologies for identifying terrorism and CSEA content is an evolving field.

1.7 This document also summarises our draft guidance to the providers of Part 3 services regarding how Ofcom proposes to exercise its Technology Notice functions. That full draft guidance can be found in Annex 5.

Our proposals for minimum standards of accuracy

We have developed an approach that could be used to set minimum standards of accuracy for terrorism/CSEA content identification technologies. We propose that the minimum standards of accuracy would be scores that technologies must meet or exceed, after a specific assessment conducted by Ofcom or a nominated third party. This consultation sets out our proposals for those assessments, and the minimum standards that must then be met.

Our proposed approach is to first, and in every case, assess the accuracy of technologies through an audit-based assessment. We have also set out an additional, supplementary stage, which is to conduct independent performance testing if technologies pass the audit-based assessment.

Ofcom is consulting on both: the audit-based assessment and supplementary independent performance testing; and whether minimum standards of accuracy should be comprised of the audit-based assessment only or include the additional, supplementary stage of independent performance testing.

Audit-based assessment

The audit-based assessment has been developed around four principles, which are the basis for assessing accuracy: technical performance, fairness, robustness, and maintainability. We have chosen these principles based on our research, which suggests these factors affect the accuracy of a technology from its development through to deployment.

For technologies to be accredited through the audit-based assessment, we would require technology developers applying for accreditation to provide evidence of their technology meeting a set of objectives that sit beneath each of the four principles. This evidence would then

be audited and scored independently against set criteria. The objectives and minimum score for passing the assessment against these objectives would be set by the Secretary of State and would function as minimum standards of accuracy. The technology would then be accredited if it meets or exceeds the minimum standards set by the Secretary of State. We provide more detail on the principles, objectives, and our rationale in Section 4.

Independent performance testing

An additional, supplementary stage to accreditation could require technologies that have passed the audit-based assessment to undergo independent performance testing.

For technologies to be accredited through this testing, they would need to meet or exceed thresholds set against selected performance metrics. We propose that the thresholds are set based on the relative performance of technologies that apply for accreditation, so that the minimum standards of accuracy would reflect current market capabilities and the highest performing technologies would be accredited. Technologies would be tested in specific categories, based on technologies that address the same harm type and that use the same data type.

The Secretary of State would not therefore prescribe minimum performance thresholds, but instead how Ofcom should calculate the applicable thresholds. The thresholds would be published and periodically updated to reflect the accuracy of any new technologies submitted for accreditation. We provide more information on this 'benchmarking' approach in Section 4 of this consultation.

Draft guidance on Technology Notices

This consultation also includes draft guidance for the providers of Part 3 services about how Ofcom proposes to exercise its Technology Notice functions when it is able to do so. As noted above, this would be after the Secretary of State has approved and published minimum standards of accuracy and, where relevant, technology has been accredited as meeting those standards.

The draft guidance sets out the process we would typically expect to follow if we consider that a Technology Notice may be necessary and proportionate, including how we would expect to engage with the provider of the Part 3 service in question.

Next steps

1.8 We are inviting stakeholders' views on our proposals for minimum standards of accuracy and our draft guidance on Technology Notices. The full list of consultation questions is in Annex 4. The deadline for responses is **5pm on Monday 10 March 2025.**

1.9 During 2025 we also plan to undertake further work to determine how the accreditation scheme will work, which may include commissioning further external research. Taking account of this work and subject to consultation responses, we will then issue our advice on minimum standards of accuracy to the Secretary of State. We expect to publish our final guidance for providers of Part 3 services on the exercise of our Technology Notice functions at the same time as our advice to the Secretary of State.

2. Introduction

2.1 This consultation is about Ofcom's new power in section 121 of the Online Safety Act 2023 ('the Act'). It focuses on:

a) our proposals for minimum standards of accuracy in the detection of terrorism and CSEA content against which technology needs to be accredited before it can be required as part of a Technology Notice. Consultation responses will inform the advice that we must provide to the Secretary of State on minimum standards of accuracy, who must then approve and publish standards before we are able to use this new power; and
b) draft guidance for the providers of Part 3 services on how we propose to exercise our Technology Notice functions when we are able to do so.

2.2 This introduction explains:

a) Ofcom's powers to tackle illegal harms and specifically terrorism and CSEA content, including our powers under section 121 and their position within the wider regulatory regime;
b) the steps required and other considerations for Ofcom, before we can issue a Technology Notice; and
c) the scope of this consultation and the Technology Notice power.

Ofcom's powers to tackle terrorism and CSEA content

2.3 The Act provides for a new regulatory framework which has the general purpose of making the use of regulated internet services safer for individuals in the UK. To achieve this, the Act imposes duties which require providers to identify, mitigate and manage the risks of harm from illegal content and activity that is harmful to children, as well as conferring new functions and powers on Ofcom.

2.4 The Act places a range of new duties on all providers of Part 3 services in relation to illegal content. The concept of 'illegal content' is discussed in more detail below. These duties differ depending on whether the service is a user-to-user or search service, and whether the content is 'priority' illegal content or relevant non-priority illegal content. They can however broadly be broken down into two categories:

a) duties to assess risks of harm arising on the service, otherwise referred to as the 'risk assessment duties'; and
b) duties to manage and mitigate those harms, otherwise referred to as the 'illegal content safety duties'.

2.5 A provider's illegal content safety duties will vary depending on whether they are providing a user-to-user service or a search service. For user-to-user services, these duties include:

a) to take or use proportionate measures relating to the design or operation of the service to prevent individuals from encountering priority illegal content and minimising the length of time that such content is present on the service;[3]

[3] Sections 10(2)(a) and 10(3)(a) of the Act.

b) to take or use proportionate measures relating to the design or operation of the service to effectively mitigate and manage the risks of harm to individuals, as identified in the service provider's most recent illegal content risk assessment;[4] and
c) to operate the service using proportionate systems and processes designed to swiftly take down (priority or non-priority) illegal content when they become aware of it. This is frequently referred to as the 'takedown duty'.[5]

2.6 For regulated search services, these duties include:

a) to take or use proportionate systems and processes to effectively mitigate and manage the risks of harm to individuals as identified in a service's most recent illegal content risk assessment;[6] and
b) to operate a service using proportionate systems and processes to minimise the risk of individuals encountering search content that is priority illegal content and other illegal content that the provider knows about.[7]

2.7 Part 7 of the Act sets out Ofcom's powers and duties in relation to regulated services. These include a specific power under section 121 of the Act to issue 'notices to deal with' two specific types of illegal content – terrorism and/or CSEA content – where we consider it necessary and proportionate. These notices are the focus of this consultation document.

2.5 We provide more detail below about what constitutes terrorism and CSEA content. We also explain Ofcom's powers in relation to this content and how Technology Notices complement our other powers. A more detailed summary of the relevant legal framework applicable to Technology Notices is set out in Annex 6.

What are terrorism and CSEA content?

2.6 Terrorism and CSEA content are both categories of 'priority illegal content' under the Act.

2.7 'Illegal content' is a new concept created by the Act, defined as 'content that amounts to a relevant offence'.[8] Section 192 of the Act sets out how, where they are required to do so, providers of services should make judgements as to whether content is illegal content. The approach set out in the Act is such that 'illegal content judgements' are to be made if the service provider has 'reasonable grounds to infer' that the content in question amounts to a relevant offence.[9] 'Reasonable grounds to infer' is not a criminal threshold, and there are no criminal implications for the user if their content is judged to be illegal content against this threshold.[10]

[4] Section 10(2)(c) of the Act.
[5] Section 10(3)(b) of the Act.
[6] Section 27(2) of the Act.
[7] Sections 27(3)(a) and 27(3)(b) of the Act.
[8] Content may consist of 'certain words, images, speech or sounds'. A full definition of illegal content may be found in section 59 of the Act. Section 59(3) of the Act sets out when content 'amounts' to an offence.
[9] The service must make this judgement using all 'relevant information that is reasonably available' to it. These two principles are more fully explained in our Illegal Content Judgements Guidance ('the ICJG'). This guidance is designed to help providers better understand what illegal content is and how they should make judgements about that content.
[10] The provider is not obliged to report illegal content to law enforcement except where the content in question is subject to requirements to report Child Sexual Exploitation and Abuse (CSEA) material to the National Crime Agency (NCA) in the UK, as set out in section 66 of the Act.

2.8 The Act sets out the 'relevant offences' in scope of the criminal law in the UK for the purposes of identifying 'illegal content'. Under the Act, the relevant offences comprise:

a) a list of priority offences, and
b) 'non-priority' (or 'other') offences.

2.9 In total there are over 130 priority offences in scope of the Act. These are set out in Schedules 5 (Terrorism offences), 6 (CSEA offences) and 7 (Priority offences) of the Act and are the most serious offences covered by the Act, as defined by Parliament. All providers of Part 3 services will need to act to prevent users encountering content amounting to one of these offences.

2.10 Terrorism content refers to content which amounts to an offence specified in Schedule 5 of the Act. These offences include, but are not limited to:

a) A series of offences relating to 'proscribed organisations'
b) Offences related to information likely to be of use to a terrorist
c) Offences relating to training for terrorism
d) Other offences involving encouraging terrorism or disseminating terrorist materials
e) Miscellaneous, more specific terrorism offences
f) Offences relating to financing terrorism

2.11 Concern around encountering terrorism content is high, with four in five (80%) UK internet users in Ofcom's Online Experiences Tracker expressing a high level of concern about content encouraging extremism, radicalisation or terrorism online.[11] Research also suggests that there has been an increase in the number of young people arrested for terrorism offences in recent years, with more than half (58%) of the 236 terrorism-related arrests for those under 18 since 2001 coming since April 2015.[12]

2.12 CSEA content refers to content which amounts to an offence specified in Schedule 6 of the Act. These offences include, but are not limited to:

a) Offences relating to the making, showing, distributing or possessing of an indecent image or film of a child
b) An offence of possession of a prohibited image of a child
c) Linking to or directing a user to child sexual abuse material (CSAM)
d) An offence of possession of a paedophile manual
e) An offence of publishing an obscene article
f) Sexual activity offences (potential victim under 16)
g) Adult to child offences (potential victim under 16)
h) 'Arranging' together with 'assisting', 'encouraging' and 'conspiring' offences which could take place between adults and/or children (potential victim(s) under 16)
i) Offences concerning the sexual exploitation of children and young people aged 17 or younger

2.13 The scale of CSEA content is difficult to ascertain due to lack of reporting and the concealed nature of offences. However, the National Crime Agency (NCA) estimates that there are

[11] Online Experiences Tracker, Wave 6 data. Question 7: '*Below is a list of things that someone may come across on the internet. Please tell me on a scale of 1 to 5, where 1 means 'mildly concerned' and 5 means 'very concerned', how concerned you are about the below existing online.'* Source: Ofcom, 2024. Online Experiences Tracker 2024. [accessed 5 December 2024]

[12] Allen, G., Burton, M. & Pratt, A., 2022. Terrorism in Great Britain: the statistics, *Commons Library Research Briefing No. CBP7612,* p. 17. [accessed 7 October 2024]

between 680,000 and 830,000 UK-based adult offenders who pose varying degrees of risk to children, equivalent to 1.3% to 1.6% of the UK adult population.[13] The Child Abuse Image Database (CAID), which retains information on child abuse imagery encountered by the police, had recorded 2.1 million unique images on its database in the year ending March 2019.[14]

Ofcom's Technology Notice powers within the wider Illegal Harms framework

Ofcom's powers in respect of Codes of Practice

2.14 Codes of Practice provide the foundation for Ofcom's implementation of the online safety regime in the UK. As required by the Act, they set out Ofcom's recommendations to regulated services about the measures they may take to comply with their new online safety duties, including their illegal content safety duties.

2.15 While service providers are not required to follow the Codes, those that do will be considered compliant with the relevant duties.[15] Services may also take what the Act calls 'alternative measures' but must keep a record of the action they take and explain how this meets the relevant safety duties.

2.16 Ofcom can include a range of measures within Codes of Practice relating to the design and operation of regulated services. These can include, but are not limited to, measures relating to regulatory compliance and risk management, the design of functionalities, algorithms and other features, policies on terms of use, user support measures and content moderation measures. The measures included within Codes of Practice are not targeted at individual regulated services. They are intended to apply either to all regulated user-to-user or search services or to specific kinds of services based on their size and capacity, and the findings of their most recent risk assessment. The Act also sets out principles that Ofcom must have regard to in preparing its Codes of Practice, including the principle that the measures included must be proportionate and technically feasible.[16]

2.17 Ofcom is also able to recommend the use of 'proactive technology' as a way of complying with some of the duties set out in the Act, including the illegal content safety duties. Proactive technology includes some kinds of content identification technology, user profiling technology and behavioural identification technology.[17] There are, however, additional constraints on Ofcom's power to include proactive technology measures in a Code of Practice. Importantly, Ofcom may not recommend the use of proactive technology to analyse user-generated content communicated privately, or metadata relating to such content.[18]

---

[13] National Crime Agency, 2023. National Strategic Assessment 2023 for Serious and Organised Crime. [accessed 7 October 2024]
[14] Office for National Statistics, 2020. Child sexual abuse in England and Wales: year ending March 2019. [accessed 7 October 2024]
[15] Section 49(1) of the Act.
[16] Paragraph 2(c) of Schedule 4 to the Act.
[17] Section 231 of the Act.
[18] Paragraph 13(4) of Schedule 4 to the Act.

Ofcom's first Illegal Content Codes of Practice

2.18 Ofcom published its Statement on Illegal Harms on 16 December 2024. This Statement contains a range of documents, including:

a) The Illegal Harms Codes of Practice; and
b) Ofcom's Illegal Harms Register of Risks, Risk Profiles, and Risk Assessment Guidance.

2.19 Our first Illegal Content Codes of Practice include a range of measures that will help make the use of internet services safer for UK individuals and reduce the prevalence and dissemination of priority illegal content including terrorism and CSEA content online. These include that the providers of regulated Part 3 services:

a) Set clear and accessible terms and conditions that explain how users will be protected from illegal content, including terrorism and CSEA content.
b) Design content moderation systems to swiftly take down illegal content of which it is aware (that may be terrorism or CSEA content). When setting prioritisation policies for content moderation systems, providers should factor in, among other things, the number of UK users encountering a particular item of illegal content and the severity of harm from that content.
c) Adequately resource and train content moderation teams to deal with terrorism and CSEA content, including to meet increases in demand caused by external events, such as crises and conflicts.
d) Have user reporting and complaints processes for illegal content that are easy to find, access and use.
e) Remove accounts if there are reasonable grounds to infer they are run by or on behalf of a terrorist organisation proscribed by the UK Government.
f) Take measures to tackle the online grooming of children, including safer default settings that make it harder for strangers to find and interact with children online.
g) Search services should take appropriate moderation action in relation to terrorism content, such as making sure this content is de-indexed or de-prioritised.
h) Provide supportive prompts and messages for child users during their online journey, to empower them to make safe choices online, such as when they turn off default settings or receive a message from a user for the first time.

2.20 The first Illegal Harms Codes of Practice also include the following proactive technology measures for certain Part 3 services:

a) Use of hash-matching technology, which automatically detects known CSAM images shared by users in their public content.
b) Use of Uniform Resource Locator (URL) detection technology, which scans public posts to remove illegal URLs that lead to material depicting the sexual abuse of children.
c) Prevention of CSAM URLs from appearing in results by search engines and applying warning messages on search services when users search for content that explicitly relates to CSAM.

2.21 These proactive technology measures to detect content, and the 'take down' measure described above in paragraph 2.19(b), share some features with our Technology Notice powers under the Act. This is because a Technology Notice may require a regulated user-to-user service to identify and swiftly take down certain types of illegal content. It is important to note that our draft guidance in Annex 5 states that we would expect to have regard to whether it is technically feasible for the service provider to meet the requirements we are considering imposing in a Technology Notice. As noted in our Illegal Harms Statement, some

services are currently unable to take down illegal content or apply our hash-matching and URL detection measures because of the way their services are technically configured.[19] To the extent this is relevant when deciding whether to issue a Technology Notice in future, we would expect to take such technical limitations into account, based on the evidence available to us at that time.

Ofcom's powers in relation to Technology Notices

2.22 Ofcom's additional powers under section 121 of the Act are intended to complement its power to recommend measures in Codes of Practice and enforce against non-compliance with the illegal content safety duties.

2.23 Under section 121, Ofcom can issue a notice requiring the provider of a particular Part 3 service to deal with terrorism and/or CSEA content where we are satisfied that it is necessary and proportionate. Such a notice can require a Part 3 service provider to:

a) use technology that has been accredited ('accredited technology'), by Ofcom or another person appointed by Ofcom, to identify and/or prevent individuals from encountering terrorism content communicated publicly; and/or
b) use accredited technology to identify and/or prevent individuals from encountering CSEA content communicated publicly or privately. A provider could alternatively be required to use best endeavours to develop or source technology to deal with such CSEA content.

2.24 Ofcom's powers in respect of Technology Notices therefore differ in some important ways from its power to recommend measures in a Code of Practice.[20] For example, a Technology Notice:

a) would impose a legal (and enforceable) requirement that a specific Part 3 service provider take specific steps to deal with terrorism/CSEA content, rather than setting out recommendations;
b) would be focused solely on the use, development or sourcing of technology to deal with terrorism/CSEA content, unlike Code measures which can relate to the wider design and operation of the service and a wider range of content;
c) where a Notice requires the use of accredited technology, would require that a specific technology that has been accredited as meeting minimum standards of accuracy be used, rather than a general type or category of technology (such as hash matching or URL detection technology);
d) could require the use of accredited technology on content communicated privately for the purposes of identifying CSEA content; and
e) could require that a service use best endeavours to develop or source technology for the purposes of detecting CSEA content communicated publicly or privately.

The process before Ofcom can issue a Technology Notice

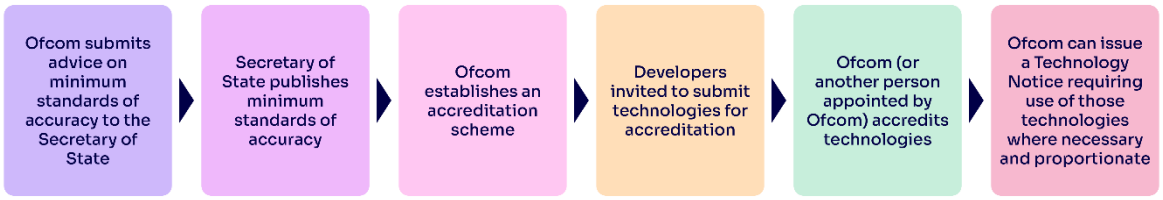
2.25 Ofcom's powers under section 121 of the Act are significant and we recognise that the use of terrorism and/or CSEA content detection technologies could impact on users' rights,

[19] See paragraphs 2.109 to 2.111 and 4.70 of Volume 2 of our Statement on Illegal Harms which explains in more detail the evidence we received from stakeholders and our own technical understanding on these points. It also explains our decisions in relation to the codes measures as a result.
[20] A more detailed summary of Ofcom's powers in respect of Technology Notices is in Annex 6.

including to freedom of expression and privacy, as well as the rights of others. We also recognise that the requirements imposed in a Technology Notice could affect the design and operation of the Part 3 service in question and impose significant costs on the provider of that service.

2.26 There are several steps that must be taken before Ofcom will be able to exercise this power. There are also some additional considerations that are built into the Act, which must be met before Ofcom will be able to issue a Technology Notice to a particular Part 3 service provider. These are set out in more detail below.

Figure 1 – a summary of the steps required before Ofcom can issue a Technology Notice



2.27 The following steps need to happen before a Technology Notice is issued:

a) **Submission of advice to Secretary of State**: Ofcom submits advice to the Secretary of State about minimum standards of accuracy in the detection of terrorism content and CSEA content. Ofcom will use responses to this consultation, alongside any further research we may commission on how accreditation could be set up, to inform our final advice to the Secretary of State.
b) **Publication of minimum standards of accuracy**: The Secretary of State needs to approve and publish minimum standards of accuracy.
c) **Establishment of the accreditation scheme:** Once minimum standards of accuracy have been approved and published, Ofcom would need to establish an accreditation scheme for specific technologies that may be used to identify terrorism and/or CSEA content. Accreditation may be conducted by Ofcom or by a third party appointed by Ofcom.
d) **Invitation to submit technologies for accreditation**: Service providers and technology developers will be invited to apply to have their technologies accredited against the minimum standards of accuracy set by the Secretary of State.
e) **Technologies are accredited:** Before we can issue a Notice requiring the use of a specific technology, Ofcom or a nominated third party must have accredited that technology against the minimum standards of accuracy.

2.28 Ofcom must also produce guidance for Part 3 service providers about how we propose to exercise our Technology Notice functions.

Considerations for Ofcom before issuing a Technology Notice to a particular Part 3 service provider (and after)

2.29 In addition to the above, there are several additional steps and considerations within the Act that Ofcom must follow in an individual case before issuing a Technology Notice to a particular Part 3 service provider. Specifically:

a) **Skilled person's report:** Ofcom must obtain a report from a skilled person to assist us in deciding whether to give a Technology Notice, and to advise about the requirements that might be imposed.[21]
b) **Warning notices:** Ofcom must give a warning notice to the service provider which outlines the requirements that Ofcom is considering imposing in a Technology Notice and provide them with the opportunity to make representations on it.[22]
c) **Necessity and proportionality:** Ofcom can only issue a Technology Notice where we are satisfied that it is necessary and proportionate to do so.[23] In determining whether it is necessary and proportionate in a particular case, Ofcom is required to consider a range of matters, including (but not limited to) the kind of service under consideration; the prevalence of, and extent of dissemination of, terrorism or CSEA content on the service; and the systems and processes already used by the service to identify and remove terrorism/CSEA content. When requiring the use of a specific technology (rather than best endeavours to develop or source technology), Ofcom is also required to consider matters such as the extent to which use of the specified technology might result in interference with users' rights to freedom of expression, and the level of risk of the use of that technology breaching privacy (including data protection) requirements.[24]

2.30 As a public authority, we must also act in accordance with our public law duties to act lawfully, rationally and fairly. Under section 6 of the Human Rights Act 1998, it is unlawful for us to act in a way which is incompatible with the European Convention on Human Rights (the ECHR). Of particular relevance to our functions under the Act are the right to freedom of expression (Article 10 ECHR) and the right to privacy (Article 8 ECHR). Other ECHR rights which may also be relevant are the right to freedom of thought, conscience and religion (Article 9 ECHR) and the right to freedom of assembly and association (Article 11 ECHR). In particular, any interference must be prescribed by or in accordance with the law, pursue a legitimate aim[25] and be necessary in a democratic society. To be 'necessary', the restriction must be proportionate to the legitimate aim pursued and correspond to a pressing social need.

2.31 Further, section 168 of the Act provides a person with a sufficient interest in the Technology Notice with the right to appeal against the Technology Notice to the Upper Tribunal.

The scope of this consultation and the Technology Notice powers

2.32 This consultation focuses primarily on the minimum standards of accuracy; our proposals for how technology could be assessed against those standards; and our draft guidance for providers of Part 3 services about how we propose to exercise our Technology Notice functions.

[21] Section 122 of the Act.
[22] Section 123 of the Act.
[23] Section 121(1) of the Act.
[24] Section 124 of the Act.
[25] As set out in Articles 8(2), 9(2), 10(2) and 11(2). The relevant legitimate aims that Ofcom acts in pursuit of in the context of our functions under the Act include the prevention of crime and disorder, public safety and the protection of health or morals, and the protection of the rights and freedoms of others.

2.33 Ofcom is also responsible for setting up a scheme to accredit technologies against the minimum standards of accuracy that must be approved and published by the Secretary of State. This consultation does not cover that scheme in detail – we focus here on standards of accuracy and how these standards could be assessed. We will say more about how we plan to set up the accreditation scheme once the Secretary of State approves and publishes the minimum standards of accuracy. However, stakeholders may wish to submit views on this, and we will continue to consider the best approach to establishing this accreditation scheme, including by undertaking additional external research where appropriate.

2.34 While we recognise the significant stakeholder interest on Ofcom's Technology Notice powers and how they might be used in future, this consultation does not take a view on:

a) The specific circumstances in which it would be necessary and proportionate to issue a Technology Notice. As explained in our draft guidance, this will be considered on a case-by-case basis in accordance with the process outlined in the draft guidance and the considerations summarised above. Our draft guidance explains that we would consider whether independent compatibility testing is appropriate before issuing a Notice, to understand the performance of the technology in the specific deployment context being considered.[26]
b) The extent to which there is technology available that could be used to identify or prevent users encountering terrorism or CSEA content in any particular deployment scenarios, for example end-to-end encrypted environments. The focus of accreditation would be solely to determine whether specific technologies meet the minimum standards of accuracy. Ofcom would subsequently be required to consider whether it would be necessary and proportionate to require them to be deployed in a particular context by issuing a Notice, as discussed in more detail in the guidance.

2.35 Ofcom has the power to require the use of technology for the purposes set out in section 121 where we consider it necessary and proportionate. The use of this power will not entail Ofcom engaging in monitoring of the content detected by that technology, whether that content is communicated via private channels or publicly.[27] Detected and unreported CSEA content present on a service will need to be reported to the NCA. This results from services' obligations to report CSEA under section 66 of the Act[28] but such content would not be reported to Ofcom.

2.36 It will remain the case that, if law enforcement wants specific information about a specific individual, to intercept communications, or to obtain communications data, then they would still need to do so using powers under (and in accordance with) the Regulation of Investigatory Powers Act 2000 and Investigatory Powers Act 2016, as appropriate.[29]

How this consultation is structured

2.37 The next three sections cover the following:

a) **Section 3 details the research and evidence that we have used to inform our proposals.** This includes the approach adopted and the themes that emerged from our research.

[26] The draft guidance is in Annex 5 – Section 3 of the draft guidance has more detail about this testing for providers of Part 3 services. We discuss the draft guidance in Section 5 of this document.
[27] Illegal Harms public/private guidance
[28] See section 66 of the Act.
[29] See Annex 6 for further detail on Ofcom's legal powers.

b) **Section 4 outlines Ofcom's proposals for setting minimum standards of accuracy**. We set out in detail our proposals for an audit-based assessment and independent performance testing, including our rationale for not setting a minimum threshold.
c) **Section 5 explains the purpose of our draft guidance for the providers of Part 3 services.** This includes how the guidance is intended to outline the typical process Ofcom would expect to follow when considering whether to issue a Technology Notice and the factors we will consider when exercising, or deciding whether to exercise, our powers.

3. Research and Evidence

3.1 This section sets out Ofcom's approach to gathering evidence and conducting research to inform the policy proposals on minimum standards of accuracy in the detection of terrorism and/or CSEA content that are set out in Section 4.

Our approach to research and evidence gathering

3.2 Over the last two years we have done the following:

a) **Desk-based research and monitoring**: a core part of our evidence gathering process is to monitor academic, industry, civil society, and journalistic research and commentary. Through this, we have developed our understanding of the technology available to identify and prevent users encountering terrorism and/or CSEA content ('**terrorism/CSEA content detection technology**'), as well as the concept of accuracy. Alongside this, we have researched how accuracy is defined and measured by other regulators, industry, and academics focused on similar technologies. We have also consulted relevant international standards, such as those from bodies like the Internal Standards Organisation (ISO) and Institute of Electrical and Electronics Engineers (IEEE).

b) **External research**: we commissioned an external consultancy, PUBLIC, in May 2023, to help us better understand how to develop an accreditation scheme. PUBLIC's research looked at how existing accreditation processes have been developed, evaluated, and operationalised from 11 accreditation approaches across five sectors.[30]

c) **Stakeholder engagement**: we spoke to a range of stakeholders, including public sector organisations, Part 3 service providers, technology developers, researchers, academics, and government organisations. There were two notable pieces of structured stakeholder evidence gathering:

i) **Multi-stakeholder Workshop:** in October 2023, we invited 30 organisations to a workshop coordinated by a market research and consulting company, Ipsos, at Ofcom's London office. The workshop helped us understand stakeholders' views on minimum standards of accuracy, the accreditation process and the Technology Notice power more generally.[31]

ii) **Information requests:** in early 2024 we issued statutory information requests to ten companies that have developed terrorism/CSEA content detection technology, including several providers of Part 3 regulated services. From the responses, we learnt how the technology developers conceptualise and assess the performance of their technologies, particularly before deployment or making them available to third parties.

3.3 This process led us to a series of research findings that we discuss in the rest of this section. They cover the following themes:

a) Understanding terrorism/CSEA content detection technologies;
b) Measuring statistical accuracy in relation to terrorism/CSEA content detection technologies;

[30] See Annex 9 for the PUBLIC Report on the Technology Accreditation Landscape.
[31] See Annex 10 for the IPSOS Multi-stakeholder Workshop Report.

c) Measuring socio-technical factors that impact accuracy in relation to terrorism/CSEA content detection technologies; and
d) Approaches to accrediting technologies against the minimum standards.

Themes emerging from our research

Understanding CSEA/terrorism content detection technologies

3.4 The Technology Notice power requires Ofcom to evaluate and accredit specific technological products, tools or solutions that can be used for terrorism/CSEA content detection, rather than broad categories of technology, such as hash-matching or keyword detection. Our research has helped us understand the different types of technologies used for this purpose and how they are deployed in regulated services. Our key findings are as follows:

a) **There is rapid growth and innovation in the trust and safety technology sector, which means the approach to setting minimum standards must be flexible enough to move with this pace of change.** This growth is being driven by increasing regulatory pressures and technological advancements, with revenues in the UK projected to reach £1 billion in annual revenues by 2026.[32] Much of this growth is due to innovations in artificial intelligence (AI) and machine learning, which are continuously evolving and enhancing capabilities in content moderation and threat detection.[33] We are likely to see this trend continue as harms evolve.
b) **There are many different types of technologies that could be used for terrorism and/or CSEA content detection.** For example, these technologies assess and classify different kinds of data (such as text, image, video, audio, metadata) and are trained on or developed from different datasets. Some technologies are designed to match data to known, previously identified terrorism and CSEA content, while others are designed to also identify and flag previously unseen terrorism and CSEA content. These technologies identify content through a variety of mechanisms, such as machine learning, hard-coded rules, and mathematical optimisation. They also change and update frequently, in line with evolving harms. The types of technologies that we expect to be considered for accreditation include but are not limited to: hashing; keyword matching; uniform resource locator (URL) detection; image-, text-, audio- or behavioural-based machine learning and AI; and rule-based technology.
c) **Although the types of technologies used to detect terrorism and/or CSEA content can be the same, they are often used differently to account for factors such as the maturity of the technology and the nuanced definitions for terrorism and CSEA.** For example, content identification technologies may be effective when dealing with CSAM content as, by its nature, CSAM content is illegal. By contrast, while a similar detection technology might also be entirely accurate at detecting properties associated with other CSEA offences such as grooming or terrorism, offences such as these rely on further analysis to prove that relevant content is in fact target content. For reasons such as this,

[32] UK Department for Science, Innovation & Technology, 2024, 8. UK Safety Tech Sector: 2024 Analysis. [accessed 24 October 2024]
[33] Etaywe, A., Macfarlane, K. and Alazab, M., 2024. Can ChatGPT flag potential terrorists? Study uses automated tools and AI to profile violent extremists, Charles Darwin University Newsroom. [accessed 24 October 2024]

the responses to our information requests revealed that technologies used to detect terrorism often worked in conjunction with human review. This is taking place in circumstances where content that has properties associated with terrorism – such as weapons in images, or as keywords in text – do not definitively mean that terrorism content is present.

d) **Technologies are often tailored to the needs of the specific service.** The UK Government's 2024 'Technology and Trust and Safety' report[34] found that services generally indicate a preference to develop content moderation tools internally, due to bespoke requirements to fit the need of the individual platform, as well as data and privacy concerns.

e) **These technologies can also perform differently based on how they are used.** For example, a hashing technology that searches for known CSAM might be expected to perform better on classes of known content than an automated content classifier (ACC), but it would be more appropriate to use an ACC when searching for unknown CSAM given that hashing algorithms are not designed for this task.

f) **Terrorism/CSEA content detection technologies usually operate within layered content moderation workflows**. Many different technologies can be used in combination to review and flag content. This was a recurring theme of the responses to our information requests and suggests that technologies of varying performance can play a vital role in more complex systems, particularly where their outputs serve as an input to other layers.

g) **Terrorism/CSEA content detection technologies are often designed to engage with and feed into human moderation systems, where human moderators trained in the identification of terrorism and/or CSEA content review, interpret and validate signals or classifications by the technology.** Participants at the Multi-stakeholder Workshop and respondents to the information requests consistently emphasised the importance of human reviewers in both ensuring oversight of the system and as an essential component of their overall terrorism/CSEA content mitigation processes.

h) **Many technologies used for the purposes of terrorism and/or CSEA detection have been developed as general-purpose technologies, which also seek to capture content which is in violation of the terms and conditions of the service.** It is possible that technologies that were not developed specifically for the purpose of identifying or preventing users' encountering terrorism and/or CSEA content may nevertheless be effective if used for this purpose and could therefore be in-scope of the Technology Notice powers. As such, a technology may meet minimum standards of accuracy, irrespective of the primary purpose for which the technology was developed.

i) **Stakeholder responses to our information requests and in discussion at the Multi-stakeholder Workshop showed the different views on the role of technology within moderation systems. Some view technology as a substitute for human moderators, while others see it as part of the triage for human review.** For example, some stakeholders suggested that when considering the accuracy of their technology, they aim for it to be as accurate as a human moderator. Other respondents made clear that they do not see technologies as a substitute for humans, nor do they expect them to achieve the same level of accuracy as humans. Such views are highly dependent on the technology in question, the task at hand, and the wider moderation system into which

[34] UK Department for Science, Innovation & Technology, 2024. Technology and Trust and Safety: The State of Play and Integration of Technology. [accessed 5 December 2024]

the technology is integrated. This shows there may be multiple ways of defining and measuring 'accuracy' depending on the technology's purpose and application.

3.5 From our review of the evidence, we understand the importance of evaluating the accuracy of technologies in specific contexts, and to fulfil specific tasks. This is because these technologies are usually only one component of a wider moderation system. We must also take a flexible approach to minimum standards so that we can adapt to the pace of change in the online safety sector.

Measuring statistical accuracy in relation to terrorism/CSEA content detection technologies

3.6 As 'accuracy' is not defined in the Act, we have considered a range of evidence to help our understanding of how it is interpreted by stakeholders and what that means for setting minimum standards. Our starting point was to consider the field of statistical accuracy - we have identified the following key findings about what it means and how we measure it:

a) **The term 'accuracy' has a specific meaning when it comes to statistical classification tasks.** In this context accuracy is defined as how close a test result and the true value of the measured phenomenon are.[35] The specific metric called 'accuracy', used to assess classification tasks, is measured by how often correct classifications are made against a labelled test dataset.

b) **However, this 'accuracy' metric is not sufficient when considering terrorism/CSEA content detection classification because it cannot by itself adequately reflect the real-world performance of a technology, particularly in the case of low-prevalence harms.**[36] To illustrate, consider a dataset with five illegal data points and 95 benign ones. If we were to take one hypothetical technology that correctly identifies 90 out of 100 of the total data points – including all five of the instances of illegal content and some false positives – that would achieve a 90% accuracy rate. Conversely, a second technology that identifies every data point in the dataset as benign – identifying no instances of illegal content – will achieve a higher accuracy rate of 95%. This is a pertinent issue for terrorism and/or CSEA content detection classification because this content makes up a small percentage of the overall amount of content online. This means technologies that cannot effectively detect terrorism and/or CSEA content could still be considered to perform well against the metric 'accuracy.'

c) **To ensure a comprehensive evaluation that captures multiple aspects of performance, robust statistical analysis of accuracy would typically include many more metrics than 'accuracy'.** These can include (but are not limited to):

   i) Precision, which measures how many data items that were predicted as 'positive' are true positives;
   ii) Recall, which measures the percentage of all true positive data items that were predicted as 'positive';[37]
   iii) F1 Score, which is a balanced way of measuring both precision and recall;

[35] International Standards Organisation, 2023. ISO 5725-1:2023(en) Accuracy (trueness and precision) of measurement methods and results. [accessed 8 October 2024]
[36] Information Commissioner's Office, 2023. Guidance on AI and data protection: What do we need to know about accuracy and statistical accuracy?. [accessed 10 December 2024]
[37] International Standards Organisation, 2021. ISO/IEC 24029 Artificial Intelligence (AI) — Assessment of the robustness of neural networks, 9. [accessed 24 October 2024]

iv) Matthews Correlation Coefficient (MCC), which is a balanced consideration of all predicted positive and predicted negative data items.[38]

d) **No single metric is a sufficiently robust measure of a technology's performance; the strength of these metrics is when measured together.** Precision and recall, for example, do not give complete information about false positives and negatives. F1 score and MCC provide a more balanced measurement of a technology's ability to classify. They can, however, be misleading for certain imbalanced data where the number of true positives is particularly small (such as with terrorism and CSEA data, as noted above).

e) **Developers of technologies may choose to optimise their technology's performance against specific metrics**. Many terrorism/CSEA content detection technologies require fine-tuning of their settings (such as model weights and parameters, sensitivity, keywords)[39] to the specific environment and use case in which they are deployed. This is usually achieved through iteration, where the deployer will start with baseline performance and then experiment by changing different parameters to affect performance across selected metrics.[40] This process is more complex for technologies underpinned by AI and machine learning, because the parameters which affect a technology's performance at identifying terrorism and/or CSEA content may not necessarily align with human intuition and expectations. Successful deployments of content detection technologies usually require ongoing support from technology experts to both achieve and maintain acceptable performance.

f) **The desired performance of a particular technology against each of these accuracy metrics may depend on the deployment context. For example, it may be more important to score highly in recall than precision, or vice-versa.** Many of the stakeholders that responded to our information requests described optimising technology for statistical precision versus recall metrics. Some stakeholders suggested that, for some technologies, the desired precision would be high. This might be the case if they relied heavily on the correct output of the technology and therefore needed to optimise for a low number of false positives. However, for other technologies, it might be preferable to prioritise recall and so have a lower precision score. This might be the case where, for example, a technology developer sought to maximise the quantity of detected content at the expense of precision and intended to complement this trade-off through extensive complementary human review capabilities to mitigate false positive results. Notably, services still considered such low-precision and high-recall technology to be 'accurate' if the technology was deployed for the purposes of detecting all positive content – despite the risk of false positive results – and where a layered workflow would serve to verify the technology's outputs.

g) **The datasets used in statistical performance testing will have a significant impact on the outcome of such testing.** There are some specific data considerations that emerged through our research. First, the data should be representative of the real-world scenarios where the technology will be applied, and the testing should be representative of the

---

[38] Also known as the phi coefficient.

[39] In Machine Learning, fine-tuning refers to techniques to further train a model whose weights have already been updated through prior training. Using the base model's previous knowledge as a starting point, fine-tuning tailors the model by training it on a smaller, task-specific dataset. (Church, K. W., Chen, Z., and Ma, Y. 2021. Emerging trends: A gentle introduction to fine-tuning. *Natural Language Engineering,* 2) [accessed 18 November 2024]. For other technologies, fine-tuning could entail updating keyword lists or heuristics, or adjusting the sensitivity of perceptual hash matching algorithms to fit the use case.

[40] Vidgen, B., 2022. The Future of Online Safety: A Data-Centric Approach, Centre for Emerging Technology and Security, 4. [accessed 9 October 2024]

data the technology will encounter to reflect the technology's capabilities.[41] Second, if a machine learning model is tested on data that is too similar to the training data, it might perform exceptionally well on the test but fail to generalise when deployed to new, unseen data (known as 'overfitting').[42] Third, it is a challenge to ensure that each instance in test datasets is accurately labelled as positive or negative for the purposes of classification.[43] This can be challenging when, for example, looking for terrorism content, as there is no universally accepted definition of terrorism, and as such, positive and negative labels may be contested. [44]

h) **It can be difficult to rely on performance test results reported by other organisations, without a detailed understanding of the data used as part of the test. The data needs to be representative, unseen to the technology prior to testing, and correctly labelled.** Without this information, any organisation verifying the performance of a technology cannot interpret or rely upon the validity of reported statistical performance testing results. In many cases, information about many of these factors, such as data inputs, is not available.[45] This can make it difficult to assess how well a technology would perform in a new environment and limit the ability to compare performance across technologies which have been subject to different testing conditions.

i) **The results of performance testing conducted on different datasets are often incomparable.** If thresholds are set against specific statistical performance metrics in the minimum standards of accuracy, these thresholds should be associated with specific, standardised tests and test conditions, and could not be generalised to apply to any and all tests conducted by other organisations.

j) **A notable absence from the evidence reviewed by Ofcom was any precedent for an agreed accuracy standard for terrorism/CSEA identification technologies, or any other technology.** Ofcom sought views on this from industry, academics and civil society as part of our October 2023 Multi-stakeholder Workshop. There was unanimous agreement that setting minimum standards of accuracy will be a difficult task. There was also agreement that specified, non-variable, thresholds for statistical measurement may be inflexible and insufficient themselves to robustly understand a technology's accuracy.[46] One of the key findings IPSOS reported from the Multi-stakeholder Workshop was a preference among stakeholders for a principles-based, or outcomes-focused and adaptable framework, which would consider wider socio-technical factors that impact accuracy.[47]

3.7 From our research into statistical measurements of accuracy, our provisional view is that the minimum standards of accuracy should consider 'accuracy' in its widest sense, using a range of metrics which give complementary insights into the technology's performance. There are

---

[41] Clemmensen, L., Kjaergaard, D., 2023. Data Representativity for Machine Learning and AI Systems, arXiv, 1. [accessed 24 October 2024]

[42] Lopez, M., 2022. Overfitting, Model Tuning, and Evaluation of Prediction Performance, Multivariate Statistical Machine Learning Methods for Genomic Prediction, 109. [accessed 24 October 2024]

[43] Information Commissioner's Office, 2023. Guidance on AI and data protection: What do we need to know about accuracy and statistical accuracy? [accessed 10 December 2024]

[44] United Nations Office of Counterterrorism, 2021. Countering Terrorism Online with Artificial Intelligence, 21. [accessed 8 October 2024]

[45] This is in situations such as when vendors refrain from releasing information that could assist offenders in exploiting or circumventing a technology, or organisations hesitate to share sensitive business information related to their deployment of a technology.

[46] Annex 10, IPSOS Multi-stakeholder Workshop Report, p.26.

[47] Ibid.

also important factors to consider about testing conditions, particularly the representativeness of the test data, and the usefulness of self-reported performance on incomparable datasets. The minimum standards should be flexible to the range of reasons for which these technologies are deployed, and therefore the alternative performance settings that may be desirable.

Measuring socio-technical factors that impact accuracy in relation to terrorism/CSEA content detection technologies

3.8 There are also socio-technical factors to consider throughout the development cycle of the technology which will impact accuracy. These are often not represented in the results of statistical performance testing. This has been recognised by industry and academics in the UK and abroad, who have proposed principles-based frameworks for evaluating AI and related technologies:

a) The UK Government's five AI principles[48] are meant to guide and inform the responsible development and use of AI in all sectors of the economy: safety, security and robustness; appropriate transparency and explainability; fairness; accountability and governance; and contestability and redress. This framework is applicable to the entire AI sector, which is much broader than the aims of this policy, but nonetheless reinforces the appropriateness of a holistic approach when considering AI evaluation.
b) We have monitored efforts made internationally towards the standardisation of technical performance principles, such as the US National Institute of Standards and Technology's 2022 Special Publication 'Towards a Standard for Identifying and Managing Bias in Artificial Intelligence',[49] which describes an approach to measuring bias mitigations across different types of biases.
c) We consulted similar principles-based frameworks set forth by the Organisation for Economic Cooperation Development (OECD) for AI trustworthiness,[50] the Information Commissioner's Office (ICO) for data protection,[51] and EU Digital Services Act for online harms.[52]
d) We also consulted international standards, such as those developed by the International Standards Organisation (ISO) and Institute of Electrical and Electronics Engineers (IEEE).[53] These are voluntary and industry-led examples of efforts to harmonise international AI regulation.

3.9 These largely principles-based frameworks highlight that factors beyond statistical accuracy can provide a comprehensive understanding of a technology's accuracy. This also points to the limitations of statistical accuracy measured under laboratory conditions alone, which will not account for potential future inaccuracies, especially in real-world scenarios. A

[48] UK Department for Science, Innovation & Technology, 2023. A pro-innovation approach to AI regulation. [accessed 8 October 2024]
[49] US National Institute of Standards and Technology, 2022. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. [accessed 8 October 2024]
[50] Organisation for Economic Cooperation and Development (OECD) AI Policy Observatory, 2023. OECD AI Principles Overview. [accessed 8 October 2024]
[51] Information Commissioner's Office, 2023. A guide to the data protection principles, 1. [accessed 8 October 2024]
[52] European Commission, 2023. Digital Services Act Overview. [accessed 8 October 2024]
[53] See for example: ISO/IEC TR 24027: Bias in AI systems and AI aided decision making (2021), ISO/IEC TR 24029-1: Assessment of machine learning classification performance (2021), and IEEE P2976: explainable AI – achieving clarity and interoperability of AI systems design (2021).

technology might perform well in a controlled test today, but security vulnerabilities and biases could lead to decreased accuracy over time or in different use cases. This underscores the importance of adhering to key principles throughout the technology's lifecycle, such as mitigating data and system biases, and ensuring robust and consistent performance. This is particularly important in the face of security threats and maintaining the technology's reliability over time.

3.10 In summary, there are other factors beyond statistical accuracy when considering accuracy for realistic deployment conditions. In particular, from this evidence we have identified principles of fairness, robustness and maintainability as crucial to ensuring the accuracy of identification technologies. These principles help address the socio-technical factors that statistical performance testing often overlooks. Demonstrable adherence to these principles can contribute to more reliable self-reported performance results (addressing concerns such as those described above in 3.6). These principles, together with performance, form a comprehensive approach to evaluating and sustaining the accuracy of technologies beyond test conditions, which means they can remain effective over time.

Approaches to accrediting technologies against the minimum standards

3.11 Technologies will be accredited against the minimum standards of accuracy, so the design of the minimum standards should be mindful of operational considerations for accreditation, to ensure the standards are robust and implementable.

3.12 We have consulted international examples of technology assessment and evaluation, particularly those of different governmental bodies, to evaluate how other regulators have attempted to measure and evaluate the effectiveness of technologies.

3.13 First, there are two reports from the USA: The White House memorandum for Executive Branch agencies 'Advancing Governance, Innovation, and Risk Management for Agency Use of AI',[54] and the US National Telecommunications and Information Administration (NTIA) 'AI Accountability Policy Report.[55] These reports recommend independent assessment of technologies for agency use and emphasise the importance of reliable data quality and sourcing in evaluations.

3.14 The White House memorandum says that in an ideal scenario, independent evaluation conducted by someone other than the technology developer should include red-teaming, audits, and performance evaluations to verify the accuracy of material claims made about these systems. This evaluation should ideally be done in as close of an environment to the technologies' operating conditions as can be achieved.

3.15 The memorandum also explains the importance of creating withheld datasets to test models in order to minimise the risk of AI developers overfitting or specifically training their technologies to perform well on test data, which can impact the reliability of test results. The NTIA report proposes that a mixed approach to AI evaluation, which combines self-

[54] United States Executive Office of The President Office of Management and Budget, 2024. Memorandum on Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence. [accessed 8 October 2024]

[55] United States National Telecommunications and Information Administration, 2024. AI Accountability Policy Report. [accessed 8 October 2024]

assessment from technology developers with an independent assessment, is the best current approach to mitigate risk of harms.

3.16 Second, we have considered AI Verify,[56] which is an AI governance testing framework and software toolkit developed by a partnership between the Singaporean Government and major technology companies. AI Verify seeks to help developers validate the performance of AI systems against a set of internationally recognised principles through standardised tests.

3.17 This framework is a key example of creating an evaluation framework from principles. AI Verify is based on 11 AI governance principles, which are consistent with those used by the OECD, Singapore, the EU, and the UK. The provided toolkit then lays out a series of checks to evaluate those principles on an objective basis, which include both process checks and technical tests. Although the AI Verify framework was not developed specifically for online safety technologies, it is an example of how to evaluate principles, which informed our approach proposed in Section 4.

3.18 These technical assessment frameworks show the importance of mixed testing approaches, including statistical analysis and principles-based assessment. They also show that a company's self-certification, without any independent oversight, will unlikely be a robust approach to meeting minimum standards of accuracy. This is corroborated by academic literature on specific risks when evaluating machine learning models, particularly through self-reported assessment, which describes questionable practices where results can be the subject of 'contamination, cherry-picking or misreporting'.[57]

3.19 Ofcom also commissioned PUBLIC[58] to develop our evidence base on how products and services are currently being evaluated and accredited in different sectors. PUBLIC's research[59] highlighted six pillars of a robust and effective accreditation for Ofcom to consider when providing its advice to the Secretary of State on minimum standards of accuracy and designing its approach to accreditation:

a) Where possible, prioritise principles over prescriptive rules to allow flexibility.
b) Ensure adaptability to changing circumstances.
c) Enable uptake through a scalable process.
d) Reduce burden for applicants to incentivise uptake.
e) Identify required expertise and skills early.
f) Establish strong governance practices upfront.

3.20 Ofcom has taken account of the findings above when developing its policy proposals on minimum standards of accuracy and will continue to do so as we design the associated accreditation process.

How the evidence informed our advice

3.21 The evidence discussed above has led Ofcom to provisionally conclude that we must take a flexible approach to setting minimum standards to make sure they have longevity and can

[56] AI Verify Foundation, 2024. AI Verify Foundation - Building Trustworthy AI. [accessed 8 October 2024]
[57] Leech, G., Vazquez, J., Yagudin, M., Kupper, N., and Aitchison, L., 2024. Questionable practices in machine learning (pre-print), arXiv. [accessed 8 October 2024]
[58] PUBLIC describes itself as 'a digital transformation partner committed to helping the public sector turn innovative ideas into practical solutions.'
[59] See Annex 9 for the PUBLIC Report on the Technology Accreditation Landscape.

adapt to the pace of change in the online safety sector. We have some additional supporting conclusions:

a) **The types of technologies that could be used with a Technology Notice are currently deployed in a range of circumstances and for a variety of purposes. The minimum standards should reflect this.** Any assessment against minimum standards will have to take account of the effectiveness of technologies in specific contexts, and to fulfil specific tasks. This is because these technologies are usually only one component of a wider moderation system.
b) **Our provisional view is that the minimum standards of accuracy should consider statistical 'accuracy' in its widest sense, using a range of metrics which give alternative insights into the technology's performance.** There are also important factors to consider about testing conditions, particularly the representativeness of the test data, and the usefulness of self-reported performance on incomparable datasets.
c) **Our evidence suggests that statistical accuracy should not be the only assessment factor, particularly when considering accuracy for real-life deployment conditions.** A technology's ability to accurately identify terrorism and/or CSEA content depends on the environment it is deployed in, the other system components it interacts with, the composition of the data it processes, and its internal workings, which are known as their finetuned parameters.
d) **There are several factors that will impact the accuracy of a technology throughout the technology's lifecycle, such as data and system biases, the robustness of development procedures and the maintainability of the technology.** Principles-based frameworks highlight that statistical accuracy measured under laboratory conditions does not provide a holistic or comprehensive understanding of a technology's true accuracy, nor its potential for future inaccuracies, especially in real-world scenarios.
e) **Existing technical assessment frameworks show the importance of mixed testing approaches, including statistical analysis and principles-based assessment.** A technology might perform well in a controlled test today, but security vulnerabilities and biases could lead to decreased accuracy over time or in different use cases. This is particularly important in the face of security threats and maintaining the technology's reliability over time.
f) **Our research has also shown that some degree of independent evaluation is necessary as it is unlikely that a company's self-certification will be a robust approach to meeting minimum standards of accuracy.** This is corroborated by academic literature on specific risks that arise when evaluating machine learning models, particularly through self-reported assessment. This research describes questionable practices like contamination and selective presentation of favourable results.

3.22 Given these provisional conclusions based on our evidential analysis, the following section outlines our policy proposals for the Secretary of State on how to set minimum standards of accuracy.

4. Minimum Standards of Accuracy

Introduction

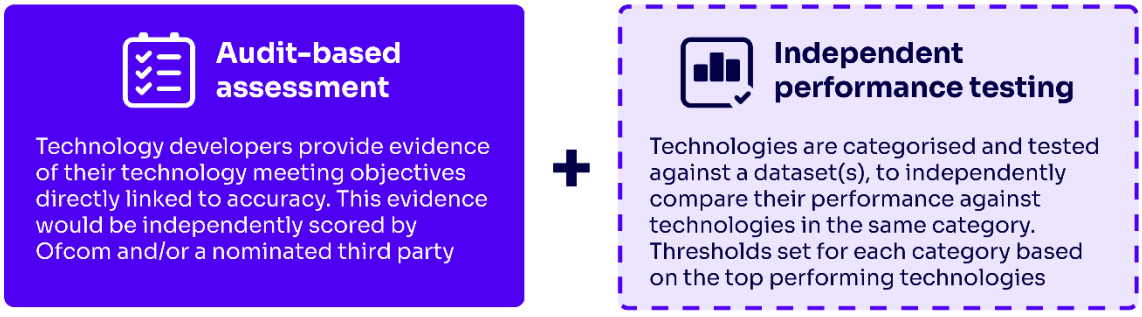
4.1 In this section, we explain our proposals for minimum standards of accuracy in the detection of terrorism and/or CSEA content and how these could function as part of an accreditation scheme.[60] We are seeking stakeholders' views on our proposals which could then form the basis of Ofcom's advice to the Secretary of State, who must ultimately approve and publish minimum standards of accuracy.

4.2 The Act does not define 'accuracy' or include any detail about how minimum standards of accuracy should be determined. In principle, there are several ways that minimum standards could be set and evaluated.

4.3 As discussed in Section 3, we have undertaken significant work – including targeted stakeholder engagement – to inform our understanding of what might be appropriate minimum standards. We have also considered the purpose of the minimum standards, which is to enable Ofcom or an appointed third party, to accredit technologies for use in a Technology Notice. Once a technology is accredited, this will not mean Ofcom will automatically be able to use it as part of a Technology Notice. As explained in Section 2, the Act includes several additional steps, including necessity and proportionality considerations, that need to be taken before Ofcom can issue a Technology Notice to a particular service provider.

4.4 With this in mind, we have developed the following policy proposal:

Figure 2 – A summary of our policy proposals



4.5 The remainder of this section is structured as follows:

a) we explain the role of proposed minimum standards and accreditation of terrorism and/or CSEA identification technologies against these standards as one step in the **high-level process** of issuing a Technology Notice, including information likely required from applicants;

[60] We focus here on standards of accuracy, but we have also considered how these standards could be assessed in practice during the accreditation process. We will say more about our approach to accreditation once the Secretary of State approves and publishes the minimum standards of accuracy.

b) we explain our proposals for an **audit-based assessment** in more detail;
c) we explain our proposals for supplementary **independent performance testing** in more detail;
d) finally, we set out our proposals regarding **re-accreditation** against the minimum standards of accuracy.

Overview of the proposed process

4.6 Before setting out the detail of our proposals regarding minimum standards of accuracy, we explain below how our proposed approach would work at a high level. In particular, we explain how an audit-based assessment and independent performance testing could form part of a two-stage process, and the wider information that we would likely request about technologies before accrediting them against the minimum standards of accuracy.

4.7 Before being considered for accreditation against the proposed minimum standards of accuracy, applicants would need to complete an **application form**, as illustrated in the template application in Annex 14. Applicants would need to provide basic information about their technology, including about which harm type(s) the technology is capable of detecting, what types of data it can process, its outputs and the compatibility of the technology. This information would not be evaluated by Ofcom or a nominated third party when determining if the technology meets the minimum standards of accuracy. Rather, it would provide information which is essential to understand the potential use cases for the technology, including potential limitations, and to categorise the technology for accreditation against the minimum standards.[61]

4.8 The application form would need to be completed in full and to the satisfaction of Ofcom before the technology could be considered against the minimum standards of accuracy.[62]

4.9 We propose that the first stage in the accreditation process would consider technologies submitted for accreditation against the **audit-based assessment**. This is discussed in more detail below.

4.10 Only those technologies that pass the audit-based assessment could thereafter be considered for supplementary **independent performance testing,** if this is included in the minimum standards of accuracy. This is discussed in more detail below.

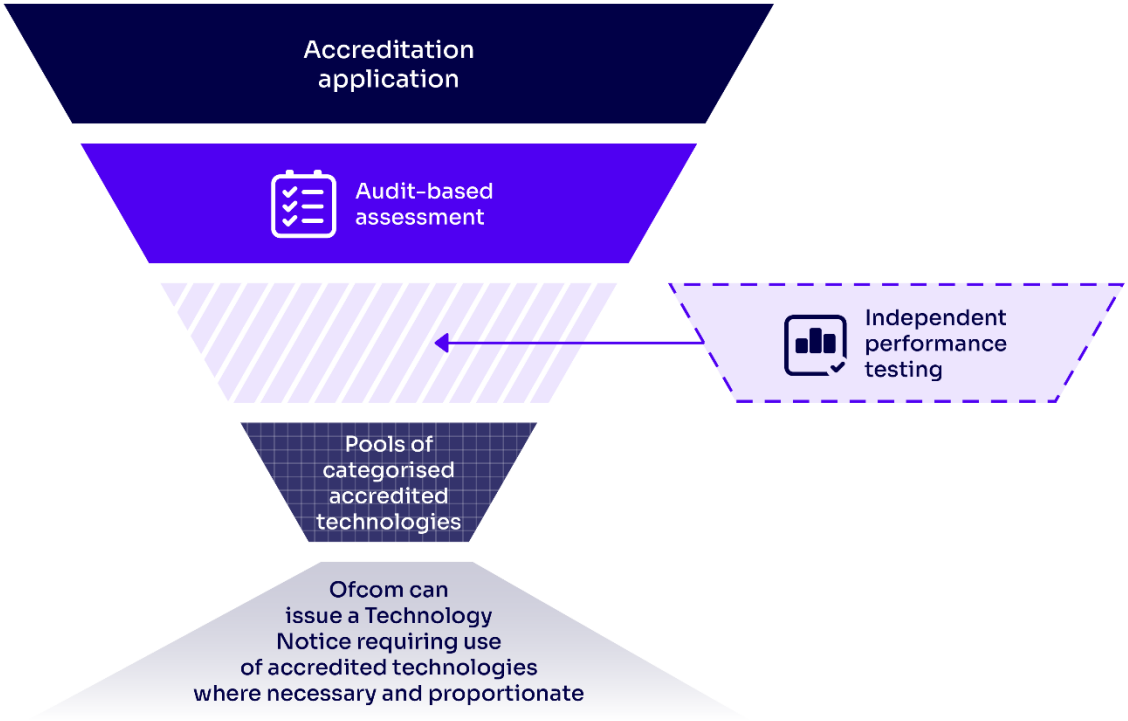
4.11 Through this approach, technologies are pooled into categories as they are assessed for their accuracy. This process acts like a funnel: through the application form and accreditation process, Ofcom would understand the potential use cases for these technologies, including the target content and intended deployment context.

4.12 Accreditation of a technology would not necessarily mean that it is necessary and proportionate for Ofcom to require the use of that technology in a Notice. Further information on this topic may be found in Ofcom's draft guidance for the providers of Part 3

[61] If the technology is ultimately accredited as meeting the minimum standards of accuracy and Ofcom is considering requiring its use in a Technology Notice, Ofcom may also use the information provided in the application form to inform its consideration of whether it is necessary and proportionate to require the use of that technology in the particular case.
[62] Ofcom would reserve the right to not consider a technology against those standards where it is found by a Court or other competent authority (such as the ICO) to have been developed in breach of UK data protection or other legal requirements.

services, set out in Annex 5. As summarised in Section 5 below, that guidance includes 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.

Figure 3 – Overview of the proposed process



Audit-based assessment

Introduction

4.13 In Section 3 we explained why the evidence shows the importance of considering several factors when assessing a technology's accuracy, and the value of a principles-based approach to technological evaluation frameworks.

4.14 In this section we outline our proposals for an audit-based assessment. The key findings from our research, as outlined in Section 3, that have led us to develop this approach to the minimum standards of accuracy include:

a) There are a broad range of technologies potentially in scope of this power. Where other bodies have attempted to codify technical assessments for a similarly broad group of technologies, they have typically opted to include principles-based assessments.
b) The optimal performance of a technology will vary depending on the deployment context. For example, developers may wish to prioritise precision over recall, or vice versa, depending on what tasks they want their technology to be optimised for. A principles-based assessment can be flexible to accommodate this.

c) Ensuring the accuracy of content detection technologies relies heavily on socio-technical factors that are often missed by statistical performance testing alone, such as bias mitigation, robustness, and maintainability.
d) While there are challenges with relying on self-reported performance data, with evidence to demonstrate adherence to such principles throughout the technology's development and performance testing, an independent assessor can interpret and verify such results.

4.15 Our provisional view is that the accuracy of terrorism and/or CSEA identification technologies should be determined by assessing technologies against the following principles:[63]

a) **Technical Performance**, which refers to the testing and reporting of a technology's ability to perform against specified metrics and technical requirements. The focus of this principle is on whether technologies can perform well at detecting terrorism and/or CSEA content in testing conditions.
b) **Fairness,** which refers to the ability of a technology to avoid unfair bias and make equal and accurate decisions across different groups of people. We recognise that some bias may be inherent. However, identifying, reporting and mitigating for, or eliminating harmful bias is essential for technology to be, and remain, accurate.
c) **Robustness,** which refers to the ability of a technology to perform reliably and maintain functionality under various conditions, including unexpected or challenging scenarios. Technology that is not robust can result in a total compromise of accuracy, as performance becomes unreliable or ineffective when the conditions under which the technology operates change.
d) **Maintainability,** which 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. This is particularly relevant for terrorism and CSEA content detection, as malicious actors frequently change tactics to bypass detection. Without maintainability, the accuracy of technology could degrade over time.

4.16 We recognise that some of the technology evaluation frameworks, which we refer to in Section 3, include other principles. However, we have sought to identify those principles that have a direct impact on the accuracy of a technology. We also note that the principles set out above are sufficiently broad to allow for other matters to be considered, even where they are often distinct principles in other evaluation frameworks. For example, we may consider security as part of the robustness principle, where relevant.

4.17 Our provisional view is that the principles above should be the basis of any assessment of the accuracy of technologies, but that they are not sufficient on their own without a robust and objective evaluation framework against which to accredit technologies. We have therefore considered how technology might be assessed against each principle and have developed the audit-based assessment as a means of doing so. We propose this could form the basis for minimum standards of accuracy.

[63] The principles proposed differ from those described in the draft Part 5 guidance consulted on by Ofcom in December 2023 and draft guidance on highly effective age assurance for Part 3 services consulted on by Ofcom in May 2024. This is due to the different technologies in scope and aims of the two policies. Ofcom is not required to accredit age assurance technologies as being highly effective.

4.18 We discuss the detail of this assessment framework below, but the key points of our proposals are as follows:

a) We have developed a set of objectives grouped under each of the four principles set out above. These are statements about how the technology has been developed, trained, or tested, and are intended to be scored against. This is similar to approaches adopted in other AI evaluation frameworks such as AI Verify.[64] We set out these proposed objectives below.
b) We intend for these objectives to be technology agnostic. They were developed to apply to the many kinds of terrorism and CSEA content detection technologies that could be accredited.
c) Any technology developer that seeks accreditation would be asked to provide evidence relevant to each of the objectives, which would then be considered by Ofcom or a nominated third party and independently assessed and scored. In line with the findings presented in Section 3, we do not consider self-certification to be sufficient.
d) We are not proposing that a technology would need to score full marks against each objective to be accredited. Rather, we propose that it should achieve a sufficient score against all of the objectives relevant to each particular principle individually (the 'principle score'), and against all of the objectives of all principles taken together (the 'total aggregated score'). We discuss our proposed approach to assessment and scoring below.
e) The audit-based assessment would not include any independent performance testing of technologies, although we would request details of self-reported performance testing and information required for Ofcom or a nominated third party to interpret the results. Our proposals for independent testing are part of our supplementary approach, which is set out in more detail below.

4.19 Both the objectives and the scoring framework would be approved and published by the Secretary of State and together would constitute the minimum standards of accuracy.

The proposed objectives

4.20 We set out below each of the objectives that we propose would form part of the audit-based assessment, and the principle to which they relate.[65]

Technical Performance

4.21 *Performance Metrics*: The technology has been comprehensively evaluated against appropriate performance metrics, and reported performance metrics along with their corresponding results are provided.

4.22 *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 generalisability to data not seen during training.

[64] AI Verify Foundation, 2024. AI Verify Foundation - Building Trustworthy AI. [accessed 8 October 2024]
[65] There are some objectives that are not relevant for technology developed without access to input/output and training data – we have labelled these (*). In these cases, we propose that an applicant seeking accreditation of such technology would not be required to provide evidence of how these objectives have been satisfied. They would instead need to confirm to Ofcom's satisfaction that their technology was developed without access to input/output and training data.

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

4.24 **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.

Fairness

4.25 **Bias Identification*: Comprehensive policies, procedures, metrics and analyses have been implemented to identify potential biases in the technology, throughout planning and development.

4.26 **Bias Mitigation:* 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.

4.27 **Data Labelling Process*: The data labelling process used for any relevant training or testing datasets is robust, with documented, standardised criteria used to label data. Measures have also been taken to ensure consistency and minimise bias and errors during the labelling process.

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

4.29 *Ongoing Bias Assessment*: A periodic assessment framework to monitor and address any emerging biases has been developed and implemented.

Robustness

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

4.31 *Consistent Performance Over Time*: The technology maintains expected operation and performance over time, demonstrating its reliability and stability when deployed in the environments it has been designed for. Any degradation over time is monitored and reported on.

4.32 *Robust Error Handling and Recovery*: The technology includes robust error handling and recovery mechanisms, enabling the management of system failures or unexpected situations.

4.33 *Reliable Operation Across Relevant Devices, system demands, and regions*: The technology operates reliably across devices it was designed to operate on, varying system capacity demands, and across different regions within the UK.

4.34 *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

4.35 *Comprehensive Documentation and Policies*: Sufficient development documentation, risk management, and data retention policies have been implemented and executed. This ensures that the performance of the technology is well-documented and managed throughout its lifecycle. Clear accountability for the documentation and management of the technology exists, and the accountable person(s) are identified.

4.36 *Proactive Risk Management*: The technology developer proactively identifies and manages risks associated with the performance of the technology at every stage of its development process through to deployment. Risk assessment plans are in place, documented, and periodically updated.

4.37 *Formal Quality Assurance (QA) Plans and Periodic Monitoring*: Formal Quality Assurance plans are in place and periodic monitoring is conducted with a view to maintaining or improving the technology's performance. The processes for updating, repairing, and improving the technology over time have been documented.

4.38 *Performance Impact Assessments:* Impact assessments are conducted before and after updating the technology to ensure that updates do not have a negative effect on performance across different devices or environments (as applicable).

4.39 *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.

4.40 *Emergency Update Procedures:* The technology provider has established procedures for handling emergency updates or repairs.

Assessment against the principles and associated objectives

4.41 Once the minimum standards of accuracy are published and the accreditation scheme established, we would publish a list of questions that correspond to each objective. This would help applicants understand what evidence to provide in support of each objective and ensure a consistent approach to scoring by the accrediting body. In Annex 11 we have included some examples of the types of questions we would expect to ask and the format in which they could be presented to applicants. Ofcom or a nominated third party would then evaluate the evidence provided in response to each of the questions and use it to determine a score for each objective.

4.42 To pass the audit-based assessment, we propose that an applicant technology would need to achieve a minimum score, rather than scoring full marks for each objective. We also do not propose that each objective should be scored on a binary pass or fail basis. Our provisional view is that this would be disproportionate and insufficiently flexible, particularly given that these are only intended to be minimum standards of accuracy.

4.43 Our proposals for how technologies would be scored, and what scores would be required to pass the audit-based assessment, are as follows:

a) **Scoring against each objective**: The evidence provided by the applicant would be independently evaluated by Ofcom or a nominated third party and a score would be determined for each objective in line with the following scoring system:

i) Five (5) points where there is robust and comprehensive evidence that the objective has been met.

ii) One (1) point where there is some, but limited evidence that the objective has been met.
iii) Zero (0) points where there is no evidence that the objective has been met.

This point distribution is designed to ensure that the scoring system reflects the importance of robust and comprehensive evidence and guarantees that if a solution only demonstrates limited evidence across all objectives, it would not be sufficient to pass the assessment.

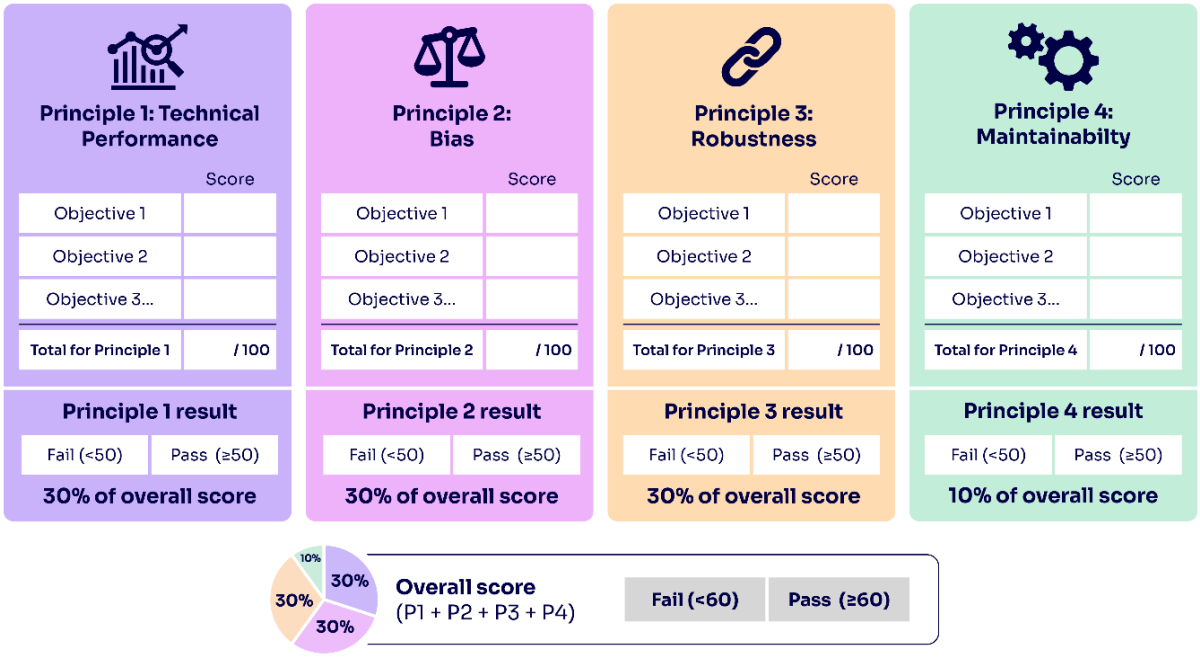
b) **Scoring against each principle**: Ofcom or a nominated third party would aggregate the scores for each objective that sits under the same principle, making a total score for that principle. We refer to this as the 'principle score'. Each of the four principle scores would then be converted into a score out of 100.[66] We propose that a technology would need to score at least 50/100 on each of the four principles to pass the audit-based assessment.

c) **Total aggregated score**: Each of the principle scores would then be weighted before being aggregated to make a total aggregated score out of 100. We propose the principles technical performance, robustness and fairness be given the same weighting (30% each) as our provisional view is that they are equally important. However, as we would expect technologies' accreditation status to be periodically reviewed (discussed below), we see maintainability as less important in this particular context and are proposing that it has a lower weighting of 10%.

We propose that a technology would also need to score at least 60/100 on the total (aggregated) score to pass the audit-based assessment. We have suggested this minimum score on the basis that this could be achieved by an applicant that provides robust and comprehensive evidence that at least half of the objectives have been met, and some (but limited) evidence that the other objectives have been met. Our provisional view is that this would be sufficient as a minimum standard of accuracy. We recognise however that this could be set higher or lower and welcome stakeholders' views on this.

[66] For each principle, the points gained from each objective would be summed up to obtain a raw score. The total raw score would then be normalised to fit within a 0 to 100 range.

Figure 4 – a summary of the assessment against the principles and associated objectives



Updating the audit-based assessment

4.44 We have designed the audit-based assessment to ensure that it can be applied to a range of different technologies. Our provisional view is that it should remain appropriate over time, even as the online safety industry grows and innovates. However, Ofcom would expect to periodically review the framework to consider whether it remains fit for purpose or whether any changes may be appropriate, such as adding or removing objectives. In this instance, we may provide further advice to the Secretary of State.

4.45 We would also expect technologies to be periodically re-accredited against the minimum standards of accuracy. We discuss this in further detail from paragraph 4.92.

A simplified illustration of what the Secretary of State could publish as the minimum standards of accuracy

It is a minimum standard of accuracy that the technology achieves:

- **a principle score of at least 50 out of 100 for each principle; and**
- **a total aggregated score of at least 60 out of 100 against the following objectives.**

The principles and objectives are:

- **Technical Performance: [list objectives]**
- **Fairness: [list objectives]**
- **Robustness: [list objectives]**
- **Maintainability: [list objectives]**

Scoring should be based on the following framework:

- **the evidence provided by the applicant should be independently evaluated by Ofcom or a nominated third party and a score determined for each objective as follows:**
  - **five (5) points where there is robust and comprehensive evidence that the objective has been met.**
  - **one (1) point where there is some, but limited evidence that the objective has been met.**
  - **zero (0) points where there is no evidence that the objective has been met.**
- **the total aggregated score should be determined by applying the following weighting to the principle scores: Technical Performance (30%), Robustness (30%), Fairness (30%) and Maintainability (10%).**

Independent Performance Testing

Introduction

4.46 Our provisional view is that the audit-based assessment may be sufficient on its own as the minimum standards of accuracy. If a technology passes that assessment, the technology developer would have shown evidence that their technology has been evaluated under suitable testing conditions. There would also be evidence of mitigation of bias in their development and testing procedures, while also demonstrating reliable performance despite variable conditions and that the technology's performance is maintainable.

4.47 As explained in Section 2, there are also several considerations that would apply before Ofcom could require the use of accredited technology in a Technology Notice. In particular, accreditation against the minimum standards of accuracy does not mean that it would be necessary and proportionate to require the use of that technology in a particular case. Before we require its use in a Technology Notice (and as explained in our draft guidance in Annex 5) we would consider whether independent compatibility testing is appropriate to understand the performance of the technology in the specific deployment context being considered.

4.48 We note that there are significant costs and operational challenges associated with independent performance testing, which Ofcom would have to consider in detail when setting up the accreditation scheme, if it were part of the minimum standards. The complexity of testing combined with the commercial sensitivity of technologies could also deter technology developers from applying for accreditation. We will take this into account when preparing our advice to the Secretary of State.

4.49 We do, however, recognise that there may be value in independent testing of technology as part of the accreditation to evaluate the technologies which have applied for accreditation and mitigate the risk of relying on self-reported data. Independent performance testing would allow Ofcom to quantitatively assess the capabilities of different technologies for the same harm against the same representative datasets and evaluation conditions. This would give further assurance on the capability and performance of each technology, as well as providing information that might inform any subsequent consideration of whether it is necessary and proportionate to require the use of any of those technologies in a Technology Notice.

4.50 Given this, we have considered a supplementary independent performance testing stage, as part of the minimum standards, for technologies that have successfully passed the audit-based assessment. Specifically, we have considered two different approaches to setting minimum standards against which independent performance testing could be undertaken:

a) **Prescribed thresholds, approved and published by the Secretary of State**. These thresholds would be predetermined numerical thresholds published by the Secretary of State, such as, for example, 95% precision and 75% recall. These thresholds would be based on performance levels deemed acceptable for the illegal content in question and would not necessarily reflect the current capabilities of technologies submitted for accreditation.

b) **Benchmarked thresholds, calculated, published, and periodically updated by Ofcom using a mechanism approved and published by the Secretary of State**. These thresholds would be based on the performance of similar technologies under identical conditions, using the same tests, metrics, and data. These thresholds would offer a realistic reflection of the current capabilities of available technologies in the market and act as a reference point for improvements and innovation.

4.51 Below, we explain our provisional view that prescribed thresholds would not be appropriate as minimum standards of accuracy in this case. We then explain our view that benchmarked thresholds could add value as additional minimum standards of accuracy, in addition to the audit-based assessment, and set out our proposals on how these could be set in practice.

Prescribed thresholds

4.52 We recognise that some stakeholders[67] may expect minimum standards of accuracy to include one or more prescribed thresholds. These would be based on wider considerations of what the minimum acceptable level of performance is, taking into account the context in which the technology is intended to be used, and the impact that the technology might have

[67] This was a view expressed by some stakeholders at the Multi-stakeholder Workshop, summarised by the following line from the Ipsos report: 'There was some concern that minimum standards may be compromised if they are developed so as to accommodate specific emerging technologies, rather than focussing on acceptable and unacceptable outcomes.' [See Annex 10, p.5]

on users' rights to freedom of expression and privacy. There are several reasons why we do not think this approach would be appropriate or proportionate in this particular case.

4.53 First, at the point of accrediting a technology, we would not yet know how the technology would be deployed if it were required by a Technology Notice; we would need this information to make a sound judgement on acceptable prescribed thresholds. As noted above, technologies may be optimised for different performance standards, and different performance standards may be preferable depending on the use case and systems in which the technology is deployed. This may be because the technology performs differently when used as part of wider systems and processes – for example, when deployed alongside other technologies and/or with human oversight. We would like to avoid setting a minimum threshold that unnecessarily and inappropriately constrains our ability to issue a Technology Notice in the future. We would therefore prefer to avoid setting thresholds that could rule out all but a narrow range of technologies at the accreditation stage, leaving us without an appropriate range of options when considering whether to issue a Technology Notice.

4.54 Second, we are concerned that it would be unnecessarily complex if the Secretary of State were to attempt to set prescribed thresholds for every potential combination of in-scope technology, harm for which that technology was deployed, and deployment context. This could require a potentially large number of prescribed thresholds to be set. Attendees at the Multi-stakeholder Workshop described a concern with prescribed thresholds that we have accounted for in the development of our proposals: 'setting thresholds that apply ubiquitously across all technology types could also stifle innovation by excluding more bespoke solutions that are designed to serve very specific functions in a specific way, possibly in combination with other tools. For example, a high threshold set to specifically protect users against risk associated with monitoring private communications may be disproportionate for a technology that only targets publicly shared content.'[68] This suggests minimum standards could be unnecessarily complex if we were to seek to apply a minimum threshold based on a judgement of acceptable performance. There are several considerations we would have to account for, such as how we would determine the different thresholds for different types or harm, or content communicated privately compared with publicly.

4.55 Third, prescribed thresholds may not be as robust as some might expect because setting thresholds in the abstract, without reference to a specific dataset(s), can be misleading. For example, a technology might achieve 75% precision on *Dataset A*, but much lower precision on *Dataset B*. Therefore, the dataset used is crucial in setting any threshold; the dataset determines what the technology is accurate at doing, and its composition will significantly impact the reported performance results for different technologies. We do not yet know how technologies are going to perform against the datasets that would be used for an independent testing stage of accreditation.

4.56 Finally, because many of these technologies have never been tested in comparable conditions, there does not yet exist a consensus on what an appropriate minimum performance level would be. Responses to our information requests showed there is no consensus on such performance standards, nor on whether technology developers use quantitative floors in their own performance assessments. Setting a threshold without such a consensus presents a risk of setting a minimum standard so high that no technologies are

[68] See Annex 10 for the Multi-stakeholder Workshop Report

accredited, meaning Ofcom cannot use the Technology Notice powers, or that the standards are not representative of existing technical capabilities, thereby potentially dampening innovation and allowing for even relatively poorly performing technologies to be accredited.

Benchmarked thresholds

4.57 Our provisional view is that there could be value in setting a performance threshold for technologies to meet or exceed by benchmarking technologies' performance against other similar technologies. This would ensure that technologies are only accredited where they are the best-performing technologies of those tested, based on independent testing.

4.58 It is also a flexible approach that accommodates the pace of change in this sector, and the technical landscape. As the relative performance of technologies changes over time, the benchmarked threshold would adapt with the market. This would negate the need for the Secretary of State to approve and publish new minimum standards of accuracy, in contrast to prescribed thresholds that would need to be updated regularly to reflect technological progress. Benchmarked thresholds should also avoid the other potential complexities presented by prescribed thresholds which we have discussed above.

4.59 We are therefore consulting on the inclusion of a supplementary stage to the minimum standards of accuracy and accreditation process, for those technologies which have passed the audit-based assessment.

4.60 We discuss our proposals in more detail in the remainder of this section. However, broadly speaking, this would involve Ofcom or a nominated third party independently testing technologies against the same dataset(s) and against thresholds set by reference to the performance of similar technologies. To be accredited, a technology would need to meet or exceed thresholds calculated by Ofcom in accordance with a 'mechanism' approved and published by the Secretary of State. This means that the Secretary of State would not specify prescribed thresholds, but instead specify how Ofcom should calculate the applicable threshold(s).

4.61 We explain below our proposals regarding:

a) the testing categories for the purposes of benchmarking technologies; and
b) the approach to testing, including: proposed thresholds and passing requirements; setting the first thresholds; and proposals for how those thresholds could be updated.

Proposed testing categories

4.62 An important consideration for the benchmarking process is how technologies are grouped. We refer to these groups as the '**testing categories**'. There can be more than one way in which technologies can be grouped and, in such a case, the choice may have important practical implications for the results of any benchmarking.

4.63 Specifically, the number of groups and the assignment of technologies to each one will be an important factor to the benchmarking process, and the calculation of appropriate thresholds. With this in mind, we have carefully considered what testing categories may be appropriate in this case.

4.64 Our provisional view is that the inclusion of technologies that detect terrorism content and CSEA content in the same testing category would not be appropriate. As explained in Section 3, we have found that although the technologies used for terrorism and CSEA content identification may be the same or similar, the content they aim to identify means that they

are often used differently and therefore optimised for distinct performance settings. We also do not consider that it would be appropriate to place technologies that analyse different types of content in the same testing category; for example, one technology that detects image-based CSEA content and another that detects text-based CSEA content.

4.65 We are instead proposing that technologies be placed into testing categories based on the following:

a) whether they identify terrorism or CSEA content; and
b) what data they analyse (for example, hashes, images, text).

We would expect this categorisation convention to be included in the minimum standards of accuracy published by the Secretary of State.

4.66 Ofcom would subsequently expect to determine the testing categories, and likely prioritise certain data types in the first instance due to the availability of test data and an understanding of technologies on the market that are likely to be submitted for accreditation. In the first instance, it is likely that hashes, images, text, URLs and video data inputs would be prioritised. We are therefore proposing the following ten priority testing categories:

CSEA	Terrorism
1. Hashes	6. Hashes
2. Images	7. Images
3. Text	8. Text
4. URLs	9. URLs
5. Video	10. Video

4.67 We acknowledge that some technologies may also process various forms of metadata and user data. It is our provisional view that our proposed approach to the minimum standards of accuracy may also apply for metadata and user data, but for the time being we are not proposing to prioritise these categories.

4.68 We recognise that our proposed approach to testing categories may result in the benchmarking of technologies whose focus and/or specific purpose may be slightly different. Both terrorism and CSEA content relate to a range of offences rather than one specific offence, and our proposed approach may not be as suitable for highly specialised technologies that rely only on specific attributes of the data.

4.69 However, our provisional view is that this approach would be appropriate and proportionate for independent performance testing as part of minimum standards of accuracy. It would be suitably high-level to capture the general capability of the technology in detecting terrorism or CSEA content (as appropriate). Within the test datasets there would be sub-categories of terrorism and/or CSEA data (see further data considerations Ofcom will consider for accreditation in Annex 13).

Proposed approach to testing

4.70 We propose to group each technology that passes the audit-based assessment into the testing categories set out above. Each technology in the same testing category would then

be independently tested by Ofcom or a nominated third party against the same dataset(s), and the same range of performance metrics would be calculated.

4.71 Through this testing:

a) Ofcom could determine the thresholds that should be in place for a particular period – these would be based on a mechanism approved and published by the Secretary of State as part of the minimum standards of accuracy; and
b) Ofcom or a nominated third party could determine how each technology submitted for accreditation performs against the thresholds in place at the time of submission, and which of these technologies should be accredited.

4.72 The full list of metrics we are proposing should be calculated as part of this independent performance testing can be found at Annex 12. Each metric would be relevant to all categories. We propose this list of metrics based on the evidence outlined in Section 3. According to this evidence, when assessing the accuracy of a technology, it is important to gather multiple metrics for a comprehensive evaluation. Metrics such as precision and recall offer an assessment of the accuracy of positive predictions, and the ability to identify all relevant instances, while the F1 score balances the trade-off between precision and recall. A more holistic understanding can be achieved by further considering metrics such as throughput and latency, which suggest how the technology might perform in environments of different scales. It is important to consider these different perspectives on a technology's performance to understand its accuracy.

4.73 The independent testing that would take place to set the threshold and assess submitted technologies would require the procurement and maintenance of datasets containing terrorism and CSEA content, as well as benign content. This brings with it a range of considerations about the type of data, the method of identification and collection, and quality assurance. If independent performance testing forms part of the minimum standards of accuracy, we would explore these issues in more depth prior to operationalising the accreditation process. We have, however, provided some illustrative considerations in Annex 13.

Proposed thresholds and passing requirements

4.74 We propose that the thresholds are set in accordance with a mechanism approved and published by the Secretary of State.

4.75 For each testing category, a threshold for each of the metrics set out in Annex 12 would be calculated based on the results of performance testing of technologies in that testing category. In practice, this would result in multiple thresholds for each testing category. Unlike the prescribed thresholds discussed above, these would be based on the measured capabilities of technologies within that testing category. Each threshold would be published along with the list of technologies accredited against them in Ofcom's annual report, which is required under section 128 of the Act.

4.76 As discussed below, these thresholds would then be updated at set periods to reflect the accuracy of technologies subsequently submitted for accreditation. Updated thresholds would also reflect any significant changes to datasets used during independent testing, as may be needed to reflect evolving harms and wider socio-technical developments, which might impact the threshold values.

4.77 There are several ways in which the thresholds could be set, but we have decided to consult on two options:

a) **Mechanism A: Threshold set at the 75th percentile of all submitted technologies during the 'previous testing period' which have passed the audit-based assessment**. This would mean that the results would be aggregated for each metric and the 75th percentile taken as the cutoff point for each. Only technologies which scored higher than the cutoff on F1 score and at least one other metric would pass. We explain what we mean by the 'previous testing period' from paragraph 4.80 below.
b) **Mechanism B: Threshold set at the 90th percentile of the top-performing technology of the 'previous testing period' which has passed the audit-based assessment.** Under this mechanism, only technologies performing no less than 10% worse than the top-performing technology of the previous testing period on F1 score and at least one other metric would pass.

4.78 We have proposed an approach based on percentiles to allow for relative comparisons between technologies designed to operate on the same data types and address similar harms. Percentiles allow for the comparison of technologies in a more granular way than averages since they are less affected by outliers. We recognise that, regardless of the mechanism and the mathematical functions underlying it, there is always a risk that a few poorly performing technologies may be accredited or that some high-performing technologies might narrowly miss the threshold(s). However, this also offers an opportunity for thresholds to serve as targets for technology developers striving to improve their performance and innovate.

4.79 There will be multiple thresholds for each testing category, as listed above. To pass the supplementary independent performance testing stage, within their testing category, we propose that the technology would need to meet or exceed the thresholds set against the F1 score and at least one other of the listed metrics. Our reasoning for this is as follows:

a) The F1 score gives equal weight to both precision and recall. It is particularly useful as a means of balancing the trade-off between precision and recall, especially when it is not clear whether the minimisation of false positives or false negatives should be prioritised. By consolidating both precision and recall, the F1 score provides a single metric that reflects both the model's ability to correctly identify positive instances while minimizing false positives and false negatives. This makes it a valuable tool for evaluating the performance of classification models.
b) However, it is our provisional view that F1 score alone is not enough. Our provisional view is that technology should have high performance in at least one other area – such as having particularly high precision, or capacity for high throughput, or low latency – which complements the F1 score and underpins an assessment of accuracy in another dimension.
c) Our provisional view is that it would be disproportionate to require technologies to meet or exceed the threshold on all metrics. This is because technologies are unlikely to be strong in all areas, as strength in one area will likely come at the cost of strength in another. For example, precision and recall usually have an inverse relationship; it is usually impossible to optimise for both. For the reasons outlined in Section 3, we believe that this would unhelpfully narrow the pool of technologies accredited and available to Ofcom for the purposes of issuing a Technology Notice, before the ultimate use case for the technology is clear.

The 'previous testing period'

4.80 We propose that thresholds are set based on the results of independent performance testing during the 'previous testing period', to reflect the capabilities of the technologies

that have most recently been submitted for accreditation. This has the advantage of enabling Ofcom to be transparent about the precise thresholds which are in force before applicants submit their technology for accreditation.

4.81 We propose that the 'previous testing period' should ordinarily be the period since the last thresholds were set. This ensures that only the results of recent independent performance testing are taken into account when setting the thresholds and that more outdated testing results do not distort the thresholds. For the first and second thresholds, we propose a slightly different approach, as explained below.

First thresholds

4.82 If the minimum standards feature independent performance testing, we would be unable to set the first thresholds in line with the description above, having not received any applications for accreditation. There would not at that stage have been any prior independent performance testing or any prior thresholds.

4.83 Instead, we propose that the first thresholds are set based on the performance of the first round of technologies that pass the audit-based assessment. Once the technologies have been tested in their testing categories, the mechanism for setting the benchmarked thresholds would kick in. The technologies would be benchmarked against each other, which would lead to the first thresholds being calculated for each testing group, and technologies being accredited accordingly. The 'previous testing period' for the purpose of the first thresholds would therefore be the period immediately before the first thresholds are set.

4.84 While this means that the first round of applicants for accreditation will not know what the threshold for accreditation is, this trade-off avoids the pitfalls of a prescribed threshold outlined above, in favour of ensuring that the thresholds would be achievable by at least some applicants passing the audit-based assessment. This would allow Ofcom to accredit the highest-performing technologies submitted for accreditation. Once the first thresholds are set, updates to the thresholds would be determined in advance, thereby providing applicants with greater clarity.

Second thresholds

4.85 For the second thresholds, we propose that the 'previous testing period' should include any testing done in the four-year period since the first thresholds were set, in accordance with the general approach set out in paragraph 4.81. However, if this approach were taken, it would not allow Ofcom to take account of the independent performance testing conducted before the first thresholds were set, as this testing would have occurred immediately before the four-year period began. We do not think it makes sense to exclude this testing. It would be only slightly more than four years old, and therefore comparable in age with testing we would be able to take into account in setting the subsequent thresholds. Accordingly, we propose that the 'previous testing period' for the purposes of setting the second thresholds also includes the period immediately before the first thresholds are set.

4.86 The second thresholds would therefore be based on:

a) the performance of the first round of technologies that pass the audit-based assessment, which would have also been used to determine the first thresholds; and
b) any testing done since those thresholds were set.

Illustrative examples of how Mechanism A and B would apply in practice

4.87 Below, we have presented illustrative examples of how Mechanism A and B would apply in practice based on performance testing of a group of hypothetical technologies.

4.88 **Table 1 models how we propose to set the first thresholds using the mechanism set by the Secretary of State, and how technologies would be accredited against those first thresholds**. It illustrates how we would use the testing scores of seven technologies in one testing category to set thresholds across five example metrics. To explain the table further:

a) Under Mechanism A, technologies '*Delta*' and '*Eta*' would be accredited.
b) Under Mechanism B, technologies '*Delta', 'Epsilon'* and *'Eta'* would be accredited.

4.89 In tables 1 and 3, [A] denotes success against the relevant threshold calculated in accordance with Mechanism A (i.e., 75th percentile of all technologies) and [B] denotes success against the relevant threshold calculated in accordance with Mechanism B (i.e., 90th percentile of top technology).

Table 1 – Dummy testing scores for the first applications for accreditation in a particular testing category.

Technologies submitted for first accreditation period	Accuracy metrics					
	F1 score	+	Precision	Recall	False Positive Rate	Latency
Alpha	58.67		44.00	**88.00** (A)	25.00	**120.00** (A,B)
Beta	61.75		**90.00** (A,B)	47.00	**7.00** (A,B)	250.00
Gamma	56.28		53.00	60.00	18.00	210.00
Delta	**75.23** (A,B)		71.00	80.00	11.00	**125.00** (A,B)
Epsilon	**73.66** (B)		59.00	**98.00** (A,B)	20.00	**130.00** (A,B)
Zeta	62.78		77.00	53.00	13.00	300.00
Eta	**75.13** (A,B)		**89.00** (A,B)	65.00	**7.00** (A,B)	180.00

4.90 **Table 2 shows the first thresholds set under Mechanism A and B**, based on the results of the dummy performance testing set out in Table 1.

Table 2 – First dummy thresholds set under Mechanism A and B

	F1 score		Precision	Recall	False Positive rate	Latency
Mechanism A	**74.39**		**84**	**83**	**9**	**127.5**
Mechanism B	**67.71**		**81**	**88.2**	**7.7**	**132**

4.91 **Table 3 models how the performance testing used to determine the first thresholds would be used to inform the second thresholds, and a new group of dummy technologies seeking accreditation against the second thresholds would be assessed**. These technologies would

need to meet or exceed the thresholds set for the F1 score and at least one other metric in the previous testing period (i.e., the period since the first thresholds were set, and immediately before). The threshold for the next (third) period of accreditation would be updated based on the performance of all the technologies tested in this second period. To explain the table further:

a) Under Mechanism A, technologies '*Theta*' and '*Kappa*' would be accredited.
b) Under Mechanism B, technologies '*Theta*', '*Iota*' and '*Lambda*' would be accredited.

Table 3 – Dummy testing scores against dummy second thresholds.*

Technologies submitted for second accreditation period	Accuracy metrics					
	F1 score	+	Precision	Recall	False Positive Rate	Latency
Theta	**75.10** (A,B)		**81.00** (B)	70.00	**8.00** (A)	155.43
Iota	**68.28** (B)		**90.00** (A,B)	55.00	**5.00** (A,B)	198.12
Kappa	**76.96** (A,B)		69.00	87.00 (A)	17.00	254.20
Lambda	**67.88** (B)		75.00	62.00	13.00	**110.52** (A,B)
Mu	62.75		59.00	67.00	21.00	**114.09** (A,B)
Nu	60.73		**98.00** (A,B)	44.00	**4.00** (A,B)	**91.98** (A,B)
Xi	65.62		**92.00** (A,B)	51.00	**6.00** (A,B)	350.29
Omicron	67.39		**87.00** (A,B)	55.00	**8.00** (A)	**129.11** (B)

*This Table 3 assumes that no further testing is done between the first and second thresholds being set (and therefore that the first and second thresholds are the same). If further testing were done, we note that first and second thresholds could be different. See paragraphs 4.85 and 4.86.

Updating the thresholds

4.92 In our proposed approach, the Secretary of State would approve and publish, as part of the minimum standards of accuracy,[69] a mechanism to set the benchmarked thresholds and this mechanism would remain unchanged.[70] The mechanism would be used by Ofcom to calculate thresholds, which would be published and periodically updated by Ofcom in line with the mechanism published by the Secretary of State. While this approach ensures an evidence-based threshold is set based on the performance of independently tested technologies, the benefits of our proposed approach would only be fully realised if the thresholds set by that mechanism were updated over time to reflect evolving technological

[69] As explained above, the minimum standards of accuracy would also comprise the principles, objectives and scoring framework for the audit-based assessment.
[70] The Secretary of State would retain the right to approve and publish new or modified minimum standards of accuracy should they consider it appropriate, following advice from Ofcom.

performance. We have therefore carefully considered how often the thresholds should be updated.

4.93 There is a balance to be struck when deciding how often to update the thresholds set by the mechanism. On the one hand, updates that are too frequent could be impractical as they will mean a high resource burden for both Ofcom and the technology developers seeking accreditation. Updates could also have significant practical implications should Ofcom require the use of accredited technology in a Technology Notice, given that such a Notice can require the use of accredited technology for up to three years. On the other hand, if updates are infrequent, the thresholds would not keep pace with technological change, therefore undermining the benefit of the mechanism that determines the thresholds and unable to capitalise on technological developments in a timely manner.

4.94 Having carefully considered this trade-off, we propose that Ofcom update the thresholds based on the results of independent performance testing in the 'previous testing period' every four years.[71][72]

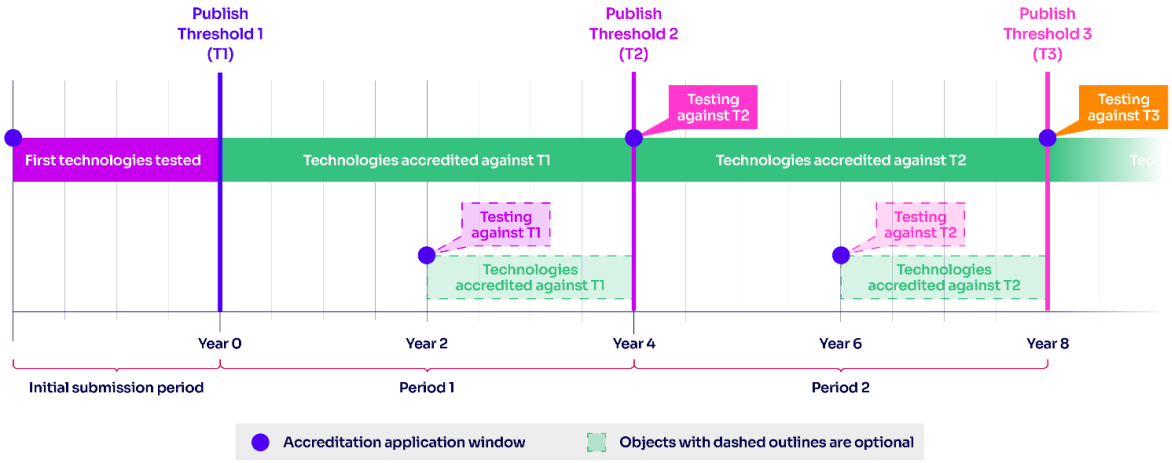
4.95 We recognise that this may result in a longer time period between updates than some stakeholders may prefer. If no further testing were conducted between the first thresholds being set and them being updated, our proposals would mean that the first and second thresholds (which would apply for a combined eight years) would be the same. However, we believe that this would still strike an appropriate balance between the considerations set out above. It would also be more appropriate than predetermined numerical thresholds specified by the Secretary of State, which would remain in place indefinitely unless and until the Secretary of State were to change them following further advice from Ofcom.

4.96 While we are proposing that the thresholds should only be updated every four years, this would not necessarily mean that technology developers could only apply for accreditation every four years. As explained in Section 2, this consultation does not cover the accreditation process in detail. The responsibility for setting up the accreditation scheme would rest with Ofcom and would not form part of the minimum standards of accuracy approved and published by the Secretary of State. However, our current expectation is that we would accept applications for accreditation during a standardised application window, rather than on an ad hoc basis. It might be appropriate to open this window every two years. We discuss the implications of this for re-accreditation below.

[71] As explained in Annex 13, Ofcom may consider it appropriate to make changes to the dataset(s) against which technologies are tested from time to time, to reflect the evolving technology and harms. Where those changes are significant, we recognise that it would not be appropriate to rely on thresholds set based on the results of independent performance testing against the previous datasets. We propose that the thresholds would be updated by re-testing the pool of existing accredited technologies (which had been accredited against the earlier dataset) against the updated dataset(s). Taking the performance results of lowest-performing technology from this pool of already accredited technologies when tested against the new dataset allows us to adjust the threshold to reflect the difficulty of the new test conditions, while maintaining the same real performance level. We would aim to align this with accreditation where possible, to avoid changes to the dataset mid-period and give sufficient notice of updated thresholds.

[72] Although as noted at paragraph 4.85 and 4.86, we are proposing that the update to the second thresholds be based on the performance of the first round of technologies that pass the audit-based assessment and any testing done in the period since those thresholds were set.

Figure 5 – Timeline for Independent Performance Testing



A simplified illustration of what the Secretary of State could publish as the minimum standards of accuracy

It is a minimum standard of accuracy that the technology meets or exceeds the performance thresholds set against the relevant metrics (F1 score and at least one other metric) by the technologies that pass the audit-based assessment in the relevant testing category in the previous testing period.

The performance thresholds are [A: the 75th percentile of all submitted technologies for each group; or B: the 90th percentile of the top-performing technology for each group]. The previous testing period is:

- **for the first thresholds, the period immediately before those thresholds are set;**
- **for the second thresholds, the period immediately before the first thresholds are set and the four-year period thereafter; and**
- **for all other thresholds, the four-year period since the last thresholds were set.**

Re-accreditation

4.97 We think accreditation against minimum standards should be time-limited. This is because, as outlined in Section 3, technologies undergo constant changes, and will need to adapt to the evolving landscape of illegal content.

4.98 Our proposal is for technology to be assessed through an audit-based assessment in every case. We recognise that the evidence provided by applicants for accreditation against each of the objectives may not change over time. However, we consider it would still be appropriate and proportionate to re-accredit technology against the audit-based assessment as the policies and practices used by technology developers whose technology has already been accredited may evolve over time, even if there are no changes to the minimum standards for the audit-based assessment itself. We are therefore proposing that technology developers be required to resubmit, and where necessary, update their evidence in support of the audit-based assessment. We would expect the re-accreditation process to be a lesser

burden on applicants than the first time, particularly if there have been no significant changes in their technology and their associated policies and practices.

4.99 Our proposal also includes a supplementary independent performance testing stage against thresholds which are updated every four years, based on the best-performing technologies from the results of independent performance testing that has taken place over the previous four-year period. If this were to form part of the minimum standards of accuracy ultimately approved and published by the Secretary of State, we note that, after the thresholds have been updated, accredited technology would need to be re-accredited against the relevant updated thresholds to retain its accredited status. This would mean technology could not remain accredited against the previous thresholds. If the independent performance testing stage is included in the minimum standards of accuracy, we would expect re-accreditation to take place after the updating of the thresholds. This would mean it would be a maximum of four years after the technology was previously accredited.[73]

4.100 Even if independent performance testing does not form part of the minimum standards of accuracy, our provisional view is that it would be appropriate and proportionate for re-accreditation against the audit-based assessment to take place every four years to mitigate the risk that there have been any significant developments that may affect the performance of the technology since it was previously accredited.

4.101 We therefore propose that technologies are re-accredited every four years, whether the minimum standards of accuracy include only the audit-based assessment, or also the supplementary independent performance testing stage.

Consultation question 1: Do you have any views on our audit-based assessment, including our proposed principles, objectives, and the scoring system? Please provide evidence to support your response.

Consultation question 2: Do you have any views on our proposals for independent performance testing, including the two mechanisms for setting thresholds; the approach to testing technologies in categories against particular metrics; and data considerations? Please provide evidence to support your response.

Consultation question 3: Do you have any comments on what Ofcom might consider in terms of how long technologies should be accredited for and how often technologies should be given the opportunity to apply for accreditation? Is there any further evidence we should consider?

Consultation question 4: Do you have any views on how to turn these proposals into an operational accreditation scheme, including the practicalities of submitting technology for accreditation? Is there any additional evidence that you think we should consider? Please provide any information that may be relevant.

[73] As noted above, we are considering an approach that may accept applications for accreditation every two years. In that scenario, for technologies that are accredited in an application window that opens two years after the thresholds have last been updated (and two years prior to the next threshold update), this would mean they would need to be re-accredited two years later, after the thresholds have next been updated.

5. Guidance to providers

Introduction

5.1 The Act requires Ofcom to produce and publish guidance for Part 3 service providers about how we propose to exercise our functions under Chapter 5 of Part 7 of the Act (our 'Technology Notice functions'). We must have regard to this guidance when exercising, or deciding whether to exercise, those functions, and keep the guidance under review.

5.2 Our draft guidance forms part of this consultation and is attached at Annex 5.

5.3 This chapter provides an overview of our draft guidance and highlights some useful points to be aware of when reading it.

5.4 We note, however, that our power to issue a Technology Notice is just one tool that we may use to deal with terrorism and/or CSEA content and is separate from our enforcement powers under the Act.[74] We do not need to have identified a compliance concern or open an enforcement investigation before exercising our Technology Notice functions.

5.5 When an issue comes to our attention regarding the prevalence or dissemination of terrorism and/or CSEA content on user-to-user or search services, we will conduct an initial assessment in accordance with the process outlined in our recently published Online Safety Enforcement Guidance (the 'OS Enforcement Guidance').[75] The initial assessment will explore appropriate action to take by considering all the tools available to Ofcom. This will include our power to issue a Technology Notice where the concern relates to terrorism and/or CSEA content, our enforcement powers, and our other non-regulatory tools, such as compliance remediation. We may consider enforcement action and the exercise of our Technology Notice functions at the same time.[76] In these circumstances, we will have regard to both sets of guidance and would expect to streamline our approach to these processes to minimise the burden placed on the Part 3 service provider in question.

5.6 Stakeholders are encouraged to read the guidance in full before responding to the consultation. Sections 1 and 2 of this consultation provide further information about the wider context and background to our Technology Notice functions, and a summary of the relevant legal framework and glossary are covered separately in Annexes 6 and 8.

The draft Technology Notice Guidance

Who the guidance is intended for

5.7 As required by section 127 of the Act, our draft guidance has been prepared for the providers of Part 3 services. This is because Ofcom's power to issue a Technology Notice is limited to the providers of such services.

[74] Our enforcement powers are set out in Chapter 6, of Part 7 of the Act.
[75] OS Enforcement Guidance. This guidance sets out how we will normally approach enforcement under the Act.
[76] We may also decide, where appropriate, to use one of the other tools available to Ofcom to resolve an issue, such as undertaking a period of compliance remediation – see the OS Enforcement Guidance for more information.

5.8 The draft guidance may also be of interest to others that would like to understand Ofcom's approach to the exercise of this power, and we welcome comments from other stakeholders on the draft guidance.

5.9 To the extent that service providers and technology developers are interested in understanding Ofcom's approach to accreditation, we note that our draft guidance does **not** set out our proposals regarding the process which Ofcom (or a third party appointed by Ofcom) would follow to accredit specific technologies as meeting minimum standards of accuracy.

Purpose, structure and scope of the guidance

5.10 The draft guidance explains how we propose to exercise our Technology Notice functions. In particular, it sets out:

- what a Technology Notice might require providers to do;
- 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.

5.11 The draft guidance sets out what our typical process would be, and we would be required to have regard to the guidance when deciding whether and how to exercise our Technology Notice functions. Ofcom would, however, have discretion to decide whether and how to act in a particular case, so could deviate from the typical process if we considered it appropriate to do so. In these circumstances, we would explain our rationale for taking a different approach.[77]

5.12 The draft guidance is divided into eight sections as set out below:

Section	What the Section covers
1	Overview
2	A summary of who the guidance relates to; the requirements we can impose in a Technology Notice; and the relevant legal framework, including the requirements on Ofcom before we are able to issue a Technology Notice.
3	How we would approach our assessment of whether it is necessary and proportionate to issue a Technology Notice, including the matters we must consider under the Act and other matters or considerations that might be relevant to our decision.
4	What might prompt us to initially consider exercising our Technology Notice functions, including how we might consider our power to issue a Technology Notice as part of our standard 'initial assessment' process, and the potential outcomes of an initial assessment.
5	What service providers could typically expect when we are considering issuing a Technology Notice, including how we would engage with the service provider; and our approach to information gathering during this stage (such as obtaining a skilled person's report).

[77] For example, in relation to representations on a warning notice, Ofcom considers that it will normally be able to reach a decision fairly and properly following written representations and without oral representations from the service. However, a service may, in any case, make a written request to make its representations orally to Ofcom in addition to any written representations. Ofcom will agree to such a request if it considers that an oral hearing is appropriate in view of the nature of the issue under consideration.

6	The stages of our process from deciding whether to issue a warning Notice, including giving the service provider an opportunity to make representations, to deciding whether it is necessary and proportionate to issue a Technology Notice.
7	The next steps following a Technology Notice being issued to a service provider, including reviewing their compliance with the Notice and the consequences of non-compliance.
8	How we would expect to approach the disclosure of information and publication about the exercise of our Technology Notice functions.

5.13 Below, we highlight useful information for stakeholders to be aware of when reading the guidance.

5.14 The guidance does not seek to set out in detail the circumstances when we might consider it necessary and proportionate to issue a Technology Notice, as each case will be considered on its own merits and in light of the given facts. The matters we are required to consider are set out in the Act.[78] However, in Section 3 of the draft guidance, we have also included a high-level summary of other matters Ofcom would expect to have regard to, where appropriate, when considering whether to issue a Technology Notice. This could include, for example, whether it is technically feasible for the service provider to meet the requirements we are considering imposing in a Technology Notice, taking into account the way the service is configured.

5.15 In Section 3 of the guidance, we also explain that it is more likely we would consider it necessary and proportionate to issue a Technology Notice relating to the development or sourcing of technology where a Notice to use accredited technology is not an option. This could be, for example, because there are no relevant accredited technologies or, where there are, it would not be technically feasible for any of those accredited technologies to be used on the service and/or they would not be sufficient to address the specific harm(s). In such circumstances, we would expect to take into account existing technological solutions or the state of development of any technology which could be used to identify or prevent users encountering CSEA content (even if not accredited).

5.16 The guidance also explains that, when reaching a view on whether to issue a Technology Notice and the requirements to be imposed, we would consider whether independent compatibility testing is appropriate to inform our view (in addition to the skilled person's report required by section 122 of the Act). This is notwithstanding that, in the case of a Technology Notice requiring the use of accredited technology, that technology would have already been accredited as meeting minimum standards of accuracy. In doing so, we would expect to have regard to:

a) the extent to which there is independent and robust evidence available to Ofcom about the performance of the technology in question, and the relevance of that evidence to the specific use case in question; and

b) the extent to which use of the technology would result in solely automated decision making (or conversely, use of the technology would result in content being detected that is identical to content already determined by humans to be illegal content).

5.17 If Ofcom decided that more detailed compatibility testing was appropriate, this could include testing the technology against specific metrics using bespoke datasets

[78] Section 124 of the Act specifies the matters which Ofcom must particularly consider in deciding whether it is necessary and proportionate.

representative of content the technology would expect to encounter on the service in question (for example, illegal versus benign content, image, video, text). This testing would measure the technology's capability at detecting and classifying the specific category(s) of relevant content we are concerned with. The technology may, for example, have been accredited to detect CSEA imagery generally, but we may be concerned about the prevalence of CSEA imagery of a specific age group on the service more specifically. In this case, compatibility testing done at this stage may focus on the performance of the technology at detecting CSEA content in a specific age group.

Consultation question 5: Do you have any comments on our draft Technology Notice Guidance?