Multi-Stakeholder Workshop

CSEA and Terrorism Tech Notices, Accuracy and Accreditation

Contents

1. Executive Summary

Background and context

The Online Safety Act 2023 (OSA), which became law on 26th October 2023, introduces rules for regulated services such as social media, search engines and messaging services, as well as other services that people use to share content online. Having received Royal Assent, Ofcom is officially the regulator for online safety in the UK. As part of its responsibilities, Ofcom has the power to require in-scope service providers to employ accredited technology to identify and remove Child Sexual Exploitation and Abuse (CSEA) and/or terrorism content[1], and/or prevent users from encountering such content by means of the service. In this document, we refer to these as CSEA/terrorism tech notices.

Technology will be accredited where it is accredited (by Ofcom or another person appointed by Ofcom) as meeting minimum standards of accuracy in the detection of terrorism content or CSEA content (as the case may be). Those minimum standards of accuracy will be approved and published by the Secretary of State, following advice from Ofcom.

The use of accredited technology to monitor content raises concerns about the economic, technological, legal and socio-political implications. Ofcom commissioned Ipsos UK to plan and deliver a multi-stakeholder workshop aimed at informing its advice to the Secretary of State on minimum standards of accuracy, and its approach to accreditation of technologies for CSEA/terrorism tech notices.

The multi-stakeholder engagement workshop took place in person at Ofcom's London office on the 2nd and 3rd October 2023, bringing together a group of relevant stakeholders to better understand perspectives on the use and application of technology to tackle CSEA and terrorist content.

Stakeholder reflections and discussion context

Stakeholders engaged in discussions respectfully and productively, ensuring that all views were properly heard. The resulting tone of the workshop was respectful and inclusive, and allowed for nuanced discussions throughout.

Stakeholders recognised that the polarised nature of the public debate often leaves little room for nuanced discussions, and that this may hinder understanding and progress. They welcomed the workshop as a setting for more productive deliberations.

Stakeholders did have several areas of general agreement despite their different views: they agreed that CSEA/terrorism tech notices should be a measure of last resort, the approach to accrediting technologies must be future-proof, and that there is a risk of causing further harm if accreditation and the issuing of tech notices is not conducted carefully.

[1] "CSEA content" means content that amounts to an offence specified in Schedule 6 of the OSA. "Terrorism content" means content that amounts to an offence specified in Schedule 5 of the OSA.

General considerations for use of CSEA/terrorism tech notices

Stakeholders generally agreed that collaboration between the regulator and the service in question should take place before considering the issuing of a CSEA/terrorism tech notice.

Stakeholders stressed the importance of the wider systems and processes that are applied alongside technology, such as the response action to any content detected by the technology. They had mixed views on the role of human moderation, which may improve accuracy but could also contribute to further harm (e.g. exposing moderators to harmful content) or risk human bias.

The differences between terrorism and CSEA content were highlighted by some stakeholders who suggested that the same approach may not be suitable for both types of content. Some noted that terrorism content may be harder to correctly identify without more context than certain types of CSEA content, and that false detection could undermine religious freedom of expression. Similarly, stakeholders noted that detecting grooming online (which will often take place through online conversations) may be far more complicated than other forms of CSEA content, such as child sexual abuse imagery, as it may require analysis of ongoing text-based content shared privately. Stakeholders generally agreed that there is unlikely to be a 'magic bullet' solution for the detection of all forms of CSEA and/or terrorism content, and that Ofcom would need to consider the specific harm type and application context both when accrediting technology for the purposes of CSEA/terrorism tech notices, and when requiring the use of such technology in such a notice.

There was concern about the financial impact on smaller services if they were required to implement accredited technologies. There were suggestions that flexibility should be available for such services if needed.

Stakeholders also highlighted the importance of innovation in the safety tech industry. As a result, they advocated for the financial burden on technology developers seeking accreditation to be minimised so far as possible.

Some stakeholders had strong concerns about the potential impacts of CSEA/terrorism tech notices on users' privacy, particularly the impacts of requiring accredited technology to tackle CSEA content that is communicated privately (including by means of end-to-end encrypted services). Some noted concerns about whether the issuing of CSEA/terrorism tech notices could force regulated services to break privacy regulations in other countries.

There was broad agreement that CSEA/terrorism tech notices should always seek to apply the least intrusive solution possible. This was linked to concerns about false positives, and how damaging these could be in the context of private conversations.

Significant concerns were raised by stakeholders around the accuracy of existing technologies (particularly those based on machine learning or wider artificial intelligence), and the severity of the implications associated with false positives. While there was some hope for the capabilities of technologies to improve quickly, stakeholders did not want to see trial and error through real-world application. 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.

There was hope that the OSA will be a catalyst for positive change, encouraging services to improve their safety measures and share successful strategies with others, to avoid the imposition of CSEA/terrorism tech notices.

Potential approaches to minimum standards of accuracy and accreditation

Stakeholders agreed that a combination of approaches should be used to assess the accuracy of safety technology (and, in particular, to set the minimum standards of accuracy). A popular suggestion was that a principles-based approach should be used to set the over-arching framework, with variable thresholds and process-based standards used to evidence performance. There was a sense that this approach would allow the minimum standards to evolve with developments in technology and behaviour, while still ensuring objective assessment criteria.

There was a strong feeling that technology products should not be accredited as generic tools or standalone solutions, but instead should be accredited for specific harm types and application contexts.

Operationalising the accreditation process

Stakeholders felt that up-to-date datasets which are withheld (i.e., whose contents are unknown to those seeking accreditation) are crucial to ensure robust accuracy testing as part of accreditation. However, they noted many limitations and challenges in creating and maintaining such a dataset, particularly where it needs to contain known and unknown CSEA or terrorism content.

There was suggestion of staggering the implementation of the accreditation process, to allow learning, and encourage early applications. For example, accreditation could focus initially on more established technologies that assess content communicated publicly.

While stakeholders understood that the OSA requires a technology to be accredited to be eligible for a CSEA/terrorism tech notice, some suggested that further assessment might need to be undertaken by Ofcom to consider the specific deployment context ahead of issuing such a notice. They thought this would be important to apply proper consideration to whether a technology is suitable for the service, harm type, and offender behaviour being targeted.

It was important to all stakeholders that any organisations involved in accreditation be trustworthy. There were mixed views on what types of organisations would be trustworthy to all parties, but there was some consensus that private companies may have commercial conflicts of interest. There was support for Ofcom conducting accreditation, possibly with the support of an independent third party.

Some stakeholders noted that accreditation timelines must consider the financial and wider resource burden on technology developers. They suggested that there should be a clear roadmap to accreditation, and that Ofcom should set clear expectations as to when and why the technology may be subject to re-accreditation. Stakeholders generally agreed that re-accreditation would be important to ensure that technologies continue to be sufficiently accurate.

2. Methodology

Background / context

The Online Safety Act 2023 (OSA) introduces rules for a wide range of online services including social media platforms, search engines and messaging services. These regulated services will have new duties to protect UK users by assessing the risk of harm from certain types of content (including illegal content) and taking appropriate steps to address those risks.

Section 121 of the OSA provides Ofcom, as the online safety regulator, with the power to issue a tech notice to require an in-scope service provider to employ accredited technology to identify and remove child sexual exploitation and abuse (CSEA) and/or terrorism content, and/or prevent users from encountering such content by means of the service.[2]

In particular, a notice may require a service provider to use accredited technology to:

a) identify and swiftly take down, or prevent individuals from encountering, terrorism content communicated publicly by means of the service; and/or

b) identify and swiftly take down, or prevent individuals from encountering, CSEA content communicated publicly or privately by means of the service.

Tech notices can also require a service provider to use ‘best endeavours’ to develop or source technology which achieves the purpose at b) and meets minimum standard(s) of accuracy approved and published by the Secretary of State.

However, the use of accredited technology to monitor content raises concerns about the economic, technological, legal and socio-political implications such as the impact on privacy and freedom of expression. Therefore:

- Ofcom can only require the use of technology in a CSEA/terrorism tech notice that has been accredited as meeting minimum standards of accuracy in the detection of terrorism content and/or CSEA content (as the case may be). Those standards will be approved and published by the Secretary of State, following advice from Ofcom; and

- Ofcom can only issue a CSEA/terrorism tech notice to an in-scope service provider where it satisfied that it is necessary and proportionate to do so. Several factors must be considered by Ofcom in deciding this, including the level of risk of harm, the service’s existing systems and processes, the potential for interference with freedom of expression, and the risk of violating privacy and data protection laws.

Due to the complex nature of these powers, Ofcom commissioned Ipsos UK to plan and deliver a multi-stakeholder workshop aimed at informing its advice to the Secretary of State on minimum standards of accuracy, and its approach to accreditation of technologies for CSEA and terrorism notices.

[2] As noted in the Executive Summary, technology will be accredited where it is accredited (by Ofcom or another person appointed by Ofcom) as meeting minimum standards of accuracy in the detection of terrorism content or CSEA content (as the case may be). Those minimum standards of accuracy will be approved and published by the Secretary of State, following advice from Ofcom.

Workshop objectives

The objectives of the workshop are to inform Ofcom's thinking and strategic approach to Section 121 of the Online Safety Act, providing insights to help guide Ofcom in its role of advising the Secretary of State on minimum standards of accuracy for technologies aimed at identifying and removing CSEA and Terrorist content, and/or preventing users from encountering this content.

Workshop design and rationale

The multi-stakeholder engagement workshop took place in person at Ofcom's London office on the 2nd and 3rd October 2023, bringing together a group of relevant stakeholders to better understand perspectives on the use and application of technology to tackle CSEA and terrorism content.

The list of stakeholders was developed by Ofcom and Ipsos UK and a purposive approach to sampling was deployed, with stakeholder categories created and quotas set for each of these. Ipsos advised Ofcom around the selection of stakeholders within each of the agreed categories, considering the need for range and diversity of voices within categories and across the sample. 30 stakeholders attended the workshop, recognising it was designed to allow for meaningful and efficient engagement while keeping conversations manageable and purpose driven. Of the 30, five stakeholders were only able to attend the second day.

The format of the two-day workshop was designed to maximise engagement and participation of all attendees. The dynamic form of stakeholder engagement sought to enable better assessment of the specific points of agreement and disagreement while allowing Ofcom to share initial thinking in this area with a range of stakeholders, given the significant interest from stakeholders on this issue. Ofcom brought expertise in the policy areas relevant to this issue and on the questions that were explored.

Each day had its own research question which was shared with stakeholders as the focus of the day. See below the two research questions, and the agenda for each day.

Day 1: *What considerations should Ofcom have when thinking about requiring the use of accredited technology for the purposes of tackling CSEA and terrorism?*

Timings	Agenda items
10:00 -10:55	Welcome, Introductions, scene setting
10:55 -11:45	Discussions: Concerns and hopes about the use of accredited technology
11.45-12.00	Break
12:00 – 12:40	Ofcom presentation and discussions
12:40 – 13:15	Discussion: reflections on the presentation, discussion of example scenarios.
13:15 – 14:15	Lunch

14:15 – 16:00	Discussions: Further discussion of example scenarios, discussion about appropriate interventions, effectiveness and impacts.

Day 2: *What should the minimum standard of accuracy and accreditation process look like?*

Timings	Agenda items
10:00 – 10:15	Welcome, recap, outline for the day
10:15 – 10:35	Presentation: Setting the scene on technical assessment
10:35 – 10:55	Discussion: Reflections on Day 1, and presentation.
10:55 – 11:10	Break
11:10 – 13:15	Discussion: Minimum standards of accuracy, possible frameworks for accreditation
13:15 – 14:15	Lunch
14:15 – 16:00	Discussions and panels: Trade-offs and thresholds for use of accredited technology. Operationalising the accreditation process.

Organisations in attendance

The following organisations were in attendance at the two-day workshop.

Bird & Bird
Government Communications Headquarters (GCHQ)
Information Commissioner's Office
Internet Society
Internet Watch Foundation
Moonshot
Home Office
Meta
Privacy International
Thorn
University of Cambridge
Department for Science, Innovation and Technology (DSIT)

Online Safety Tech Industry Association (OSTIA)
Open Rights Group
University of Bristol
Imperial College London
National Cyber Security Centre (NCSC)
National Society for the Prevention of Cruelty to Children (NSPCC)
TechUK
United Kingdom Accreditation Service (UKAS)
Competition and Markets Authority (CMA)
Google
National Crime Agency (NCA)

Purpose of this report

The purpose of this report is to capture the views expressed in the two-day multistakeholder engagement.

Disclaimer:

This report provides a high-level summary of the views expressed by stakeholders who were at the workshop. It is important to note that, given the limitations of in-person engagement, not all interested stakeholders were present, and therefore, this report does not capture the perspectives of all stakeholders. The purpose of this summary is to provide a condensed overview of the discussions that took place over a two-day period.

Please note that this report does not set out Ofcom's views in this area, nor does it necessarily represent any policy position that Ofcom may adopt as the online safety regulator.

Ofcom will utilise this engagement, along with other relevant evidence and wider public consultation to inform its advice to the Secretary of State on minimum standards of accuracy and its approach to accreditation.

3. Stakeholder reflections and discussion context

Chapter summary

Stakeholders engaged in discussions respectfully and productively, ensuring that everyone's views were properly heard. The resulting tone of the workshop was respectful and inclusive, and allowed for nuanced discussions throughout.

Stakeholders recognised that the polarised nature of the debate leaves little room for nuanced discussions, and that this may hinder understanding and progress. They welcomed the workshop as a setting for more productive deliberations.

Stakeholders did have several areas of general agreement despite their different views: they agreed tech notices should be a last resort measure, they agreed the approach to accrediting technologies must be future proof, and that there is a risk of causing further harm if accreditation and the issuing of tech notices is not conducted carefully.

Stakeholders acknowledged the complexities of the Online Safety Act and the implications for Ofcom.

The tone of the workshop was respectful and inclusive.

Ahead of discussions beginning, stakeholders took time together to discuss an approach to creating an environment conducive to navigating challenging discussions where opposing views were present. Stakeholders collectively defined the terms of their interactions by establishing a set of ground rules that would set the tone of the workshop.

The ground rules established by stakeholders prioritised mutual understanding over mere agreement, respectful dialogue over contention, clarity over ambiguity, and inclusivity over dominance. Stakeholders committed to listening respectfully, learning from one another, being comfortable with disagreements, and providing specific examples for clarity. They also agreed to make space for everyone's views, seek clarification when needed, celebrate areas of agreement, and set aside preconceived narratives.

The tone of the workshops was marked by these ground rules and stakeholders' respect for differing viewpoints, a commitment to understanding alternative perspectives, and a willingness to identify areas of shared values and agreement. This approach was instrumental in ensuring that stakeholders could have productive discussions, learn from each other, and identify areas of agreement and shared values, notwithstanding the contentious nature of the subject matter.

Stakeholders recognised that the polarised nature of the public debate often leaves little room for more nuanced discussions.

Stakeholders reflected on the current public and political debates around the OSA. They generally noted the polarisation of the discourse surrounding the OSA (and, in particular, section 121). One stakeholder pointed to the powers being presented as, on the one hand, an invasive surveillance tool and, on the other, a component of the OSA necessary to protect children. Stakeholders felt that this polarisation often leaves little room for productive and nuanced discussions.

Some stakeholders suggested that there are certain narratives that dominate public discourse. There was mention of the public narrative being dominated by discussions around encryption and client-side scanning. Some stakeholders felt that this debate often overlooks the fact that these concepts are not commonly interpreted or explicitly mentioned in the OSA. There were also concerns about what some stakeholders considered to be an assumption that there already exists technology to effectively tackle illegal content in encrypted environments without infringing on users' privacy.

Stakeholders further identified certain aspects of the OSA that they felt had been overlooked in the public narrative. For instance, stakeholders pointed to how the OSA also provides Ofcom with the power to require that specific companies use their 'best endeavours' to develop or source technologies to tackle CSEA content.

There was also mention of how discourse surrounding the OSA was often vague, with stakeholders pointing to different concerns they have seen raised around issues including the scope of the Act, the accreditation process, and the timeline for implementation. Stakeholders questioned what could trigger use of the powers and what it would look like in practice, for example in terms of evidentiary requirements or engagement with regulated services. These were explored through stakeholder-led discussions about how powers might be implemented in different scenarios generated by participants.

There were different levels of policy and technical literacy among stakeholders.

Throughout the discussions, it was clear that different stakeholders had differing views and understanding of some important concepts, including end-to-end-encryption and client-side scanning. Stakeholders therefore emphasised the importance of specificity when discussing these concepts to facilitate an informed understanding of their implications.

Another point of confusion was the terminology used in the Act. For example, some stakeholders felt that the term 'accreditation' might confuse some stakeholders, suggesting that what is described is closer to a 'certification scheme'. They noted that 'accreditation' is typically given to organisations or institutions whose operations meet a set of standards, whereas a technology or product would be 'certified'.

Additionally, there were mixed levels of understanding of the Act itself, with stakeholders raising various questions about the scope of Ofcom's powers, and the implications of regulatory intervention. While these questions were beyond the scope of the workshop, capturing these perspectives and areas of confusion could provide valuable insights for future clarification and guidance.

Despite differing stakeholders views, there were large areas of convergence:

- Stakeholders agreed that tech notices should be a 'last resort' measure, used when collaborative approaches to find solutions have not worked.
- Stakeholders agreed that the approach to accreditation and issuing tech notices needs to be designed so that it is as future proof as reasonably possible, given the rapidly changing landscape, especially given that it could need to accommodate technology that doesn't yet exist, or is in very early stages of development.
- Stakeholders agreed that additional harms could be caused if the approach to accreditation and issuing tech notices is not carefully considered.
- Stakeholders were divided on the extent to which users' right to privacy should be compromised for the benefit of safety.

There was an appreciation around the complexities of the OSA and the implications for Ofcom.

Stakeholders acknowledged the complexity and difficulty of Ofcom’s role in implementing the OSA. They recognised that the journey towards the OSA’s implementation would not be straightforward. However, they valued the engagement process as an opportunity to contribute their own insights and expertise.

4. Considerations for use of Tech Notices

Chapter summary

General considerations

Stakeholders generally agreed that the issuing of CSEA/terrorism tech notices should be a last resort measure, and collaboration between the regulator and regulated services should take priority. Stakeholders stressed the importance of the wider systems and processes that are applied alongside technology, such as the response action to any content detected by the technology. They had mixed views on the role of human moderation, which may improve accuracy but could also contribute to further harm or risk human bias.

The differences between terrorism and CSEA content mean the same approach may not be suitable for both types of content. Some noted that terrorism content may be harder to correctly identify without more context than certain types of CSEA content and incorrect detection could undermine religious freedom. Similarly, stakeholders noted that detecting grooming online may be far more complicated than other forms of CSEA content.

Economic considerations

There was concern about the financial impact on smaller services if they were required to implement accredited technologies. There were some suggestions of additional flexibility being available for such organisations if needed.

Stakeholders also highlighted the importance of innovation in the safety tech industry. As part of this, they advocated for the financial burden on technology developers seeking accreditation to be minimised as far as possible.

Societal considerations

Some stakeholders had strong concerns about impacts on privacy, particularly with regards to end-to-end encrypted services. There was broad agreement that tech notices should always seek to apply the least intrusive solution possible. This was linked to concerns about how damaging false positives could be in the context of private conversations.

There was hope that the OSA will be a catalyst for positive change, encouraging services to improve their safety measures and share successful strategies with others, to avoid the imposition of CSEA/terrorism tech notices.

Technological considerations

There was some concern that standards may be compromised if they are developed to accommodate specific emerging technologies, rather than focussing on acceptable and unacceptable outcomes.

There were significant concerns around the accuracy of existing technology, and the severity of implications that false positives could result in. While there was some hope for the capabilities of technologies to improve quickly, stakeholders did not want to see trial and error through real-world application.

Stakeholders felt that tech notices must consider different harm types and application contexts, seeking granular solutions rather than a ‘magic bullet’ solution.

Legal considerations

Some stakeholders had concerns about whether the issuing of tech notices could force regulated services to break privacy regulations in other countries.

Stakeholders discussed a wide array of important considerations that they felt must be taken into account when planning the use of tech notices, and how and when they should be issued. This chapter groups these considerations into general, economic, societal, technological, and legal considerations.

4.1. General considerations

The issuing of tech notices should be a last resort measure, and collaboration between the regulator and tech companies should take priority.

There was an acknowledgement that Ofcom's tech notice powers are some of the most controversial of the OSA, and as such stakeholders voiced either an expectation or an assumption that these powers will rarely be used (and will only be used if all else fails). This was an area of consensus across the room.

> **"I would expect it to be very rarely used. Where nothing else has worked and then you have to finally reach for this one."**

In the event of reports or claims of illegal content sharing, stakeholders shared an expectation that there would need to be active engagement from Ofcom and the regulated service through a two-way dialogue. This will provide Ofcom with the opportunity to assess the veracity of reports, explore what measures are already in place and how the sharing of illegal content could have come about.

There was also a sense from some stakeholders that the services that show willingness/evidence of wanting to engage and address Ofcom's concerns should be given more time, if needed, to explore alternative solutions before a tech notice is issued by Ofcom compared to those who refuse to do so. There was also recognition that where claims of illegal content being present on a service have been verified, companies may genuinely want to collaborate, but internal investigation can take time and Ofcom should recognise this. At the same time, this leeway would need to be balanced with an assessment around the immediate risk to children/people involved.

There was also consensus in the room around the importance of Ofcom validating and verifying reports or claims and collaborating with regulated services early on in the process to avoid issuing a tech notice.

Importance of the wider systems and processes that are applied alongside technology.

Stakeholders flagged the importance of the process after detection and getting this right. There was a suggestion that there should be a process which allows involved parties to defend and dispute legitimate activity and stop this from being flagged again in the future. There was reference to some highly publicised examples, where stakeholders claimed that it could be argued that the technology worked as intended but the process which followed was poorly executed, demonstrating the importance of a solid review and appeals process.

> **"It's not just the technology, it's what we're making as an accusation or process on the back of it".**

There were mixed views on the role of human moderation. There was a sense, among some, that it feels disproportionate to always escalate content review to moderators. However, other stakeholders explained that they would expect there to generally always be a role for humans in the moderation process.

> **"I think it's important to have human oversight, at least for the process to ensure quality control and that these systems are maintained".**

Stakeholders did however express a number of concerns around human moderation, as well as some of the related perceived risks:

- **A lack of consistency across moderators** depends on the company and the individuals involved

- **The risk of false negatives and false positives:** poor moderators can flag benign data as unacceptable and vice versa.

- **Additional burden on human moderators:** related to the point around the perceived inaccuracies of technologies, there was a feeling that this will bring about extra burden for moderators and result in them reviewing large volumes of content that they ideally wouldn't need to see (which links to privacy, see below).

- **The risk of cutting corners and a lack of faith in how data will be used:** a minority view was that low-cost companies involved in reviewing may cut corners and make money out of leaking data or blackmailing users.

The differences between terrorism content and CSEA content mean they need to be treated differently.

Some stakeholders highlighted that terrorism content and CSEA content are very different; there is different behaviour and patterns, and different motives for sharing such content, thus there needs to be tailored approaches for addressing these through the OSA. Some highlighted what they considered to be an absence of reliable databases for terrorism content, and the highly contextual nature of terrorism content (which can mean that, even for human moderators, it can be difficult to determine whether content should be considered as amounting to terrorism content). They suggested that there is often a fine line between terrorism content and legitimate content (including legitimate religious or political speech), and that this made them more concerned about the ability of technology (or even humans) to accurately detect such content.

There was a comparison made in one discussion around the implications of false positives for CSEA content and terrorism content. Some stakeholders emphasised that false positives with regards to terrorism content could create the possibility that individuals or groups are inappropriately criminalised, marginalised or robbed of their right to freedom of expression within the law. For this reason, these stakeholders flagged that context matters even more so when considering terrorism content.

Detecting grooming online may be far more complicated than other forms of CSEA content.

During a few of the discussions, stakeholders raised the point that developing standards and rules for online grooming content is going to be very difficult given that unlike detecting images, targeting grooming would require monitoring a conversation over time and may require more subjective judgements. Some also suggested that until databases of grooming content exist, it would not be possible to develop a standard of accuracy that can be objectively assessed, and then accredit a technology as reliably meeting that standard.

4.2. Economic considerations

There was concern about the financial impact on smaller organisations.

Stakeholders considered the implications of Ofcom's new powers for smaller firms. They flagged the increased burden on smaller businesses to meet Ofcom's requirements with regards to investigations or the issue of a tech notice. They stressed the importance of the right technology to match the harm (see section 4.4) and the right support being available to smaller organisations that do not have the infrastructure that larger organisations have.

There was also a sense that Ofcom will need to build in some flexibility in how to use its powers when working with smaller businesses, given the burden of applying technology, the resources needed (i.e., legal and compliance) and/or the levels of knowledge required.

Stakeholders saw a need to encourage improvement without stifling or discouraging innovation.

Stakeholders talked about the importance of avoiding stifling, and even disincentivising, innovation through online safety regulation. Some considered that the burden of the minimum standards of accuracy and accreditation process could curb innovation if it is too costly to make business sense for companies developing new technology.

Recognising the opportunity for innovation, some suggested having frameworks that encourage innovation to reflect the evolving playing field and the differing objectives that we might be looking to achieve through this future regime.

> **“There are things we haven’t thought of yet that could work. Where the investment goes is important”.**

4.3. Societal considerations

There was hope that the OSA will be a catalyst for positive and meaningful change.

There were discussions across the room about the potential for online safety regulation (including the minimum standards of accuracy and accreditation of technology) to ‘raise the floor’ for online safety technologies. For example, some stakeholders suggested that minimum standards could be used to make clear what is best practice in terms of the accuracy of online safety technologies, recognising that there is variation between regulated services in how they monitor and address the sharing of illegal content.

> **"There are opportunities to keep people safe... where there are good opportunities that are not being taken, and I hope this will be a catalyst. Whether it's through codes of practice, whether it's through companies wanting to get ahead of the game, whether it's through technology notices. I hope this will be a catalyst for positive change”.**

There was a sense that it is the responsibility of regulated services to design their services so that these risks and harms are very unlikely to happen in the first place. Stakeholders hoped that the OSA might encourage services to invest in improving their safety measures, and to share successful strategies with others, if they see this as a way to avoid the imposition of tech notices. This was seen as having the potential to facilitate better collaboration and knowledge sharing within the online industry.

In the context of meaningful change, some stakeholders noted that this huge and complex debate should be one that is opened up to involve a wider set of stakeholders than those in the room, such as the general public and parliamentarians.

Stakeholders expressed a preference for preventative solutions, rather than solely detection.

Given concerns raised by some stakeholders over the accuracy of available technology, the consequences of false positives, and impacts on user privacy, there was a discussion about whether detection is always the right approach, with some stakeholders cautioning that it shouldn’t be the first port of call. In the context of child grooming, the feeling here was that the focus should be on measures that can be put into place that prevent offenders from being in touch with children in the first place, which would be more effective than detecting harm after it has happened.

> **“Detection shouldn’t be the first port of call. If you can prevent an offender being in touch with a child that’s much better than detecting harm after its happened. Jumping first to detection is where we get ourselves in difficulty.”**

Concern about adverse impacts on privacy, particularly with regards to end-to-end encrypted services.

There was near consensus amongst stakeholders that tech notices should seek to apply the least intrusive method available, that would still meet the objectives (which is also a matter Ofcom is required to consider as part of its assessment of necessity and proportionality).

Early concerns were raised by civil society and industry representatives around the implication of tech notices for end-to-end encryption. This linked to a perception, among some stakeholders, that there is ambiguity in the OSA about the extent to which Ofcom might be able to require the weakening or removal of encryption in a tech notice.

> **"Particularly focusing on end-to-end encryption services. We're particularly concerned that none of these solutions can technically work without compromising the security of users"**

> **"It's important to have a truly private space that humans can interact in and removing that, I think, causes a more cyclical problem"**

The risks around intruding on privacy was a focus of discussion for some, given the precision or accuracy that can be expected from the use of automated technologies to identify content.

> **"We shouldn't be ignoring the possibility that we could reduce harm in ways that have almost no impact on intrusiveness...The idea that escalation to Moderators is not zero harm, it shouldn't be done on a whim."**

Teenagers sharing consensual nudity content was often used as an example, across discussions, of how false escalation to moderators could be a significant intrusion on the privacy of those involved, and how warning messages might be a more appropriate response.

> **“Teenager to teenager nude images, completely consensual, is highlighting it to a moderator the right thing, or is it a warning? We’re concerned about intrusion.”**

This wasn’t the only view in the room, however. Some were comfortable with the compromise around false positives as a result of detection using these technologies, if the end result is improved child safety.

> **“I know enough about how law enforcement works and the checks people go through, the security and the approach, that if the cost of protecting children is that some of my messages might be revealed to law enforcement. Even if there might be some false positives in there, I might be comfortable with that.”**

There was a fear expressed by some stakeholders around a ‘slippery slope’, with the idea that technology (including, for example, client-side scanning) could be misused and could be used by government for other purposes in the future.

4.4. Technological considerations

Concern that standards may be compromised if developed with specific emerging technologies in mind.

There was a shared concern among stakeholders around building the minimum standards of accuracy and accreditation around technology that is not yet available, or around emerging technologies that are new and unknown. For some stakeholders, it was difficult to get past the idea that the technology is not yet available, with genuine concern around how quickly it will be available and how accessible it will be for all.

There was a perception among other stakeholders that there is a risk of developing standards of accuracy on the basis of the capabilities of emerging technology, rather than starting with what is deemed unacceptable and then finding the technology that can adequately address this.

Concerns around the accuracy of existing technology, particularly machine learning.

A debate around the appropriate level of precision started on the first day and continued into the second when stakeholders discussed minimum standards of accuracy (see the next chapter 5). Stakeholders recognised that false positives can have a serious impact on people's lives, that the technology will not always be accurate, and that even low-levels of false positives can adversely impact a large number of people when applied at mass-scale.

> **"I truly struggle to see how your accuracy is going to get close enough because if the false-positive rate is higher than the prevalence of the content you're trying to detect you just get absolutely flooded."**
>
> **"Images get misunderstood, people may get reported to the police, then have to prove their innocence. Pictures taken in a private context are often innocent. These are life endangering scenarios. It can take years to build up your reputation again."**

There were stakeholders that were concerned that a 'trial and error' approach would be implemented, if minimum standards are set to accommodate current capabilities, while our current technology is further developed. Stakeholders were concerned about serious consequences if this learning curve is explored through real world application, rather than in testing environments. Due to this concern, they wanted minimum standards to focus on acceptable outcomes, not on the current capabilities of technology.

However, others were more optimistic about the likelihood of technology improving quickly (machine learning in particular), in an ever-changing landscape.

There must be considerations of different harm types and application contexts.

An area of agreement among stakeholders was that there is no 'magic bullet' in terms of the technology that will address all types of harm, across different regulated services, taking into account varying contexts.

> **"I've worked for 3 years now in this space and what we've mostly found is harm archetypes are very different, that we need very bespoke and complimentary solutions to try and tackle this harm and that content scanning is, beyond a very real privacy concern of the future, it's the least effective power … I think the concern is there is this sense of a silver bullet and the only question is, 'How do we make it happen?'…We hope that the debate will not just be solely focussed on one technology that has very real flaws, but instead embrace the complexity of the problems and the complexity of the solution."**

Stakeholders agreed on the importance of having the right accredited technology to match the harm type, having the right threshold relating to that harm, and using this at the right time. Some stakeholders highlighted that focusing the accreditation process and issuing of tech notices around the harm type, not the technology available, will lead to better targeted solutions. Another consistent theme was around the importance of being clear about the objectives and what the technology is trying to achieve.

> **“What is the harm archetype? What is our objective? What are we trying to prevent? Are we trying to prevent that material being available? Are we trying to prevent it being algorithmically promoted? Are we trying to avoid it being visible? Or are we trying to identify users and shut down groups? Because those are quite different objectives. Are we trying to prevent people being radicalised by it?”**

Some stakeholders suggested that Ofcom will need to think about harm types at quite a granular level, although others questioned the practicability of this; particularly if it would mean many different minimum standards for different types of harms and contexts.

Another point of agreement across stakeholders was the importance of context (when considering both whether to accredit technology and whether to require its use in a CSEA/terrorism tech notice). Whether this is the specific harm type, different languages, the cultural context, whether the content is known or unknown, or the users and services involved. There were examples given using age of users to demonstrate this point. One example was the difference between a site where users routinely share legal intimate photos of themselves, compared to a site such as a bird-watching site where this would be rare. A stakeholder explained that the impact of false positives (from CSAM detection technology) may be higher on a site where users routinely share intimate images, as these photos are more likely to be falsely flagged as illegal, and the impact of exposing these images as a result of false detection would specifically harm those who shared them. With this comparison in mind, they felt that even if a technology is accredited, the differing impacts and risks of requiring that technology on different services should be carefully considered by Ofcom before issuing a tech notice.

> **“It's coming back to the standard, to the matrix-based approach but it's also reacting. You were saying that the minimum standards of accuracy are separate from a lot of this context, well, I'm not sure that they can really be separate from the context. … The context's just vital in terms of what you decide to do about [content] and what the balance is between having a really positive impact on a harm, and missing the horrible things, and overwhelming other agencies or moderation teams with meaningless noise.”**

During the workshop, stakeholders generated scenarios where they discussed actions that services could take to take to address illegal content on a service:

- **Enforceable Terms and Conditions and policies:** Services setting out what behaviour is acceptable and not on the service, and what action will be taken if these terms are not adhered to. Some stakeholders flagged the downsides to too many enforceable policies, in that this can force bad actors to move onto forums with less content moderation.
- **User reporting functions:** channels for users or other stakeholders (e.g. trusted flaggers) to report content which violates services' terms and conditions.
- **Deterrence messaging:** pop up messages that warn the user about content they might be intending the share.
- **Content moderation:** systems and processes to identify and action target content. These often involve the use of automated technologies, provided by either the regulated service itself or third parties.
- **Hashing:** broadly considered to be an effective means of identifying certain types of content, i.e. known child sexual abuse material. However, hashing technologies are not infallible.
- **Age verification:** systems and processes to identify the age of users and take action on that basis, such as restricting functionalities or blocking account creation. And linked to this **network disruption** for users.
- **Methods to identify higher risk groups:** or groups that are likely to share content.
- **Safety by design features:** such as stopping users from searching or creating groups that contain certain keywords.
- **The use of cross- service intelligence:** and considering the whole ecosystem in which users operate, but essentially using tool to cross match people with other userbases where they are known to have an interest in, are reacting to or promoting the sharing of CSEA or terrorist content.
- **Using meta data machine learning**: as opposed to content machine learning.

4.5. Legal considerations

Flexibility vs specificity of the OSA and how it will work alongside other regulations.

Across discussions, some stakeholders spoke of a lack of specificity in the OSA, although there were a range of views about whether this is a challenge or a necessary feature. Some suggested that operationalisation of the CSEA/terrorism notice powers may be difficult in practice due to this perceived ambiguity – especially considering the range of services, technologies and types of harm to which a notice might relate. At the same time, there was a recognition that if the OSA was really specific, it would quickly become outdated as the technologies, behaviours and types of harm move on.

There were also some concerns about how CSEA/terrorism tech notices might interact with other regulations in various countries, including privacy regulations. For example, some attendees expressed a concern that such notices may require the use of client-side-scanning technology and that regulated services may be required to breach privacy law to comply with them (specifically, by not obtaining informed consent before deploying that technology).

5. Minimum standards of accuracy

Chapter summary

Stakeholders generally agreed that a combination of approaches should be used to assess minimum standards of accuracy. A popular suggestion was that a principles-based approach should set the framework, with variable thresholds and/or process-based standards used to evidence performance against each principle. There was a sense that this approach would allow the minimum standards to evolve with developments in technology and behaviour, while still ensuring objective assessment criteria.

There was a strong feeling that technology products should not be accredited as generic tools or standalone solutions, but instead should be accredited for specific harm types and application contexts.

Principles-based**:** This was seen as the best method for an outcomes-focused adaptable framework, but concerns were raised about whether this would be sufficiently objective by itself as an accreditation tool.

Process-based: This was seen as well suited to ensure continuous assurance, balancing objective, and measurable assessments with some flexibility to adapt to application context.

Specified (non-variable) metrics: These were viewed by some as too inflexible to accommodate the different contexts in which the technology might be used, and concerns were raised about whether these would be sufficient by themselves to robustly understand a technology's accuracy.

Variable metrics: This was viewed as promising for tailoring standards to relevant context, but could spiral into a complex landscape of standards.

A combination of models should be used to assess minimum standards of accuracy.

At the start of Day 2, stakeholders heard from two academics about potential approaches to minimum standards of accuracy and accreditation, and the challenges involved.

The discussions that followed were framed around four potential approaches to minimum standards of accuracy:

- **Principles-based:** Minimum standard based on whether a defined set of principles are adhered to in the development / deployment of a particular technology.
- **Process-based:** Minimum standard based on having rules or process in place (e.g: testing specified metrics.
- **Specified (non-variable) metrics:** Minimum standard based on assessment of a technology's performance against specified metrics (e.g., "95% true positive rate")
- **Variable standards:** Similar to specified metrics above, albeit with variable minimum standards based on, for example, type of offence or content.

The feedback on each of these approaches is outlined further into this chapter. However, there was a general agreement amongst stakeholders that no single approach would be suited, and that the best way forward is likely to be a combination of the four.

Technology products should not be accredited as generic tools or standalone solutions.

Stakeholders felt that the issuing of a tech notice should specifically target a type of harm identified as being systematic on a regulated service. As noted in the previous chapter, they spoke about there being no single technology that could feasibly be deployed as a general tool to target multiple harm types on multiple service types, while preserving privacy.

They therefore wanted to see the accreditation process test for minimum standards of accuracy against the specific type of harm and content that a technology claims to be designed to address. This was important to them so that technologies would only be accredited for the use-cases they are appropriate for, and therefore cannot be applied via a tech notice for any other context (without further accreditation).

> **“The importance of being really precise about the kinds of harm we're talking about. I think that was something we very much agreed on at this table, that you cannot have a generic technology. You need to say, 'We are looking at this particular kind of harm.' And be targeting that.”**

As part of this sentiment, stakeholders argued that technologies should be accredited for the role they play within any wider systems and processes, rather than necessarily as standalone solutions. For example, if technology is intended to be used in conjunction with human moderation, then this could be taken into account for the purposes of accreditation (and should be reflected in any tech notice requiring that technology).

Minimum standards must be designed so that they can evolve with developments in technology and behaviour.

Stakeholders often discussed whether minimum standard thresholds should represent the ‘floor’, being the basic standards that must be met as a minimum, or the ‘ceiling’, setting a higher standard for outcomes.

In discussions some stakeholders felt it would be more feasible and beneficial to, at least initially, focus on minimum standards being the ‘floor’. They argued this accommodates for the limited range of technology currently developed and incentivises companies to work towards a reasonably attainable level of performance. Stakeholders did caveat that if this approach was taken, the minimum standards should increase over time as technology develops, as well as in light of deployment experience.

> **"A minimum standard that you might compose this year might be different for a minimum standard you might compose in 5 years' time."**

Aside from threshold levels, stakeholders felt that the assessment model would have a significant impact on the ability of the accreditation process to evolve and adapt in time with developments in technology and behaviours. They often spoke about the ‘cat and mouse’ game of those seeking to share and receive CSEA and terrorism content, and methods to prevent it. For this reason, stakeholders were cautious about an overly prescriptive approach to setting minimum standards that would not allow flexibility, limiting innovation and variation in tools.

Principles-based: Preferred option for an outcomes-focused adaptable framework, but not objective enough by itself as an accreditation tool.

As described earlier in this chapter, stakeholders generally expected a combination of the proposed approaches to be important components of any minimum standards of accuracy and accreditation process. Typically, they foresaw a principle-based approach sitting above other methods, outlining the framework, with process-based and variable-threshold metrics used as methods for assessing technologies against each principle.

> **“A good start is a principles approach to evaluating what’s going on. I think it’s perfectly possible to then take some of that and start translating it into the processes that you would use to provide assurances that the principles were being adhered to. And I think that’s probably necessary from an industry point of view. Then I think that can lead to variable thresholds.”**

A key benefit stakeholders saw in a principle-based approach is its adaptability. They deemed principles could outline the outcomes that we do and do not want to achieve, which are unlikely to change, and therefore will not become quickly outdated.

> **“In cyber security we have moved towards principles based, as the outcomes sought typically don’t change, but the threat, technology, and measurements needed to assess them do.”**

Stakeholders appreciated the flexibility of how principles may be tested and evidenced, which, by not being overly prescriptive, can be tailored to the context, technology and behaviours at hand, and can be updated in line with new capabilities and industry standards. They argued that this is key to being able to keep up with the ‘cat and mouse game’ of ever evolving behaviour of those seeking to distribute or receive CSEA and terrorism content.

Another key benefit that stakeholders saw in a principles-based approach is that it may allow more variation in the technologies that achieve accreditation, which more prescriptive models may prevent. They felt that by providing different options for demonstrating evidence for each principle, technologies can be assessed in a way that considered the purpose. It may also reduce the burden of seeking accreditation for smaller companies, encouraging innovation and bespoke solutions.

However, stakeholders were clear that a principles-based approach is unlikely to be sufficient on its own, as the criteria for each principle may be ambiguous or too broad. A stakeholder explained that defined, measurable baselines are needed to make objective and consistent assessments.

> **“We cannot accredit the concept of principles because what we have is a thing called an ‘Object of conformity assessment,’ which you can measure, and then you can certify against it. So for this, the second part the process, we would expect the organisation to have a process that can be assessed, with a set of rules around it so that you can make a statement to say, ‘This organisation’s process fulfils these requirements.’”**

Stakeholders discussed how the principles could be formed, and what principles should be included in a framework.

The EU AI Act was cited as a good example of checklist principles by several stakeholders, including academics. Some stakeholders also found the REPHRAIN Safety Tech Challenge Fund evaluation criteria presented earlier in the day to be a good starting point for the development of principles.

Drawing from these examples, and their own priorities, stakeholders suggested the following principles as important to consider:

- **Transparency and accountability**: Stakeholders generally agreed that there should be a degree of transparency about how technologies work, but were divided on the extent of this transparency. A few stakeholders argued that the public and regulated services need to be able to trust these products and therefore must be able to examine them and their capabilities. These stakeholders acknowledged that transparency does have a trade-off in that particularly sophisticated bad actors may be able to assess algorithms and work around them if they are public. To mitigate this trade-off, there were suggestions that there should be full transparency in how the tool works, but training methods and data used for testing might not need to be published.

However, many felt that public availability of this information was not necessary (and carried its own risks), suggesting that the right balance would be complete transparency with an accreditor, but that the workings of the technology or algorithms should not be publicly accessible. Overall, stakeholders felt there was a need to strike a balance between maintaining transparency, protecting proprietary information, and ensuring that the technology is accountable to those who use it and those who are impacted by its use.

> **"I was uncomfortable with the idea that some of these organisations hide behind trade confidentiality when trying to evaluate their products. I do feel like there should almost be ideally complete transparency with any accreditor, even if the entire code isn't public, at least public to the accreditor so they can go in, inspect the algorithm, inspect whatever recipe was used to train the model. And maybe have some way of verifying that the model they evaluate is the same one being used by the platform and that they haven't made significant changes between evaluation and actual deployment. "**

- **Usability and maintainability:** some suggested that accredited technologies should be usable by a range of different services. This was seen as key if services may be ordered to implement tools they are unfamiliar with, especially for smaller and less-resourced organisations.

- **Fairness / lack of bias**: some stakeholders suggested that fair and unbiased technology may be hard to achieve in practice (for example, where there are inherent biases in existing data sets, or other biases such as how people tend to report content). However, stakeholders generally felt that minimum standards should seek to reduce and mitigate potential for bias as much as possible, and that transparency about the biases in any specific technology is key to this. Stakeholders also suggested that if data sets become less biased after some products have been accredited, they should then be re-assessed against those updated data sets. Some perceived a difference between bias that limits the benefits of the tool, and bias that disproportionately harms certain groups. They felt that the first is more acceptable, as it may still have a net-positive impact and can improve as advancements are made. The latter may silence or unfairly criminalise marginalised groups, particularly for terrorism content, and therefore may need different thresholds of acceptable accuracy.

- **Adaptability**: individual technologies may be at risk of becoming less effective or accurate as online behaviour and technology changes. Therefore, a key principle suggested was that the technology is able to be adapted after the point of deployment.

Process-based: Well suited to continuous assurance, balancing measurable assessments with some flexibility.

Stakeholders felt that process-based assessments and criteria are well suited to ensuring standards are built into the development and maintenance of a technology, rather than being proven in a snapshot of time.

> **“We need a suitable standard to ensure the accuracy isn't just a one-time thing, but built into the process”**

Process-based minimum standards were also seen to be fairly adaptable, so that they can accommodate changes in technology and context.

Stakeholders referenced several existing process-based accreditation schemes, which could be drawn on to shape standards for accreditation of technologies or be used as they are to accredit technologies against certain principles. Existing ISO standards in quality assurance, software development, and information security were seen as expected standards that should be factored into accreditation.

Stakeholders also suggested vulnerability disclosure programmes as a good example of the type of process-based assessment that could be appropriate to maintain minimum standards after accreditation.

> **"Is it being used in the vendor and are they reacting to it? That's actually a really good indicator as to whether or not it's going to maintain its capability."**

Some stakeholders felt that process-based standards could ensure that any changes in the design of a technology must have followed a design change process that had been validated and recorded.

Aside from the technology itself, some stakeholders suggested that the minimum standards should include process requirements for individuals who deploy or develop the technology, such as researchers or developers who will access sensitive data as part of design or deployment. These process-base minimum standards may include personnel screening, training, or even individual accreditation.

Specified (non-variable) metrics: unlikely to be sufficient to understand a technology's accuracy given different contexts.

Of the four approaches discussed, specified performance metrics were seen as the most complicated to deploy effectively. Stakeholders argued that it is unlikely that any single metric is going to be a suitable assessment across the variety of technology types, purposes, and regulated service contexts.

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 risks associated with monitoring private communications may be disproportionate for a technology that only targets publicly shared content.

> **“It risks stifling innovation, the more that it risks solutions that could be really good in one particular context, being ruled out by the fact that performance metrics have been designed with another context in mind."**

On the other hand, there were suggestions that specified metrics may give false assurance of accuracy or effectiveness by oversimplifying minimum standards to a metric, while overlooking the detail of its design, governance, and deployment.

> **"I think when you bring in the different techniques a company could use to combat terrorism, to combat CSAM, it becomes a much more principles-based, process-based, variable-threshold-based standard. I think over-indexing, maybe on the first one, actually may lead to companies not having all of the techniques to try and overall improve accuracy. "**

There were also concerns that minimum standard metrics may be difficult to test realistically as part of accreditation, as laboratory conditions are never going to fully replicate performance on live services. This was seen to be particularly concerning ahead of deployment at scale on larger regulated services.

However, some stakeholders did caveat that specified metrics may have a limited role for certain minimum standards, such as false positive rates.

Variable standards: Promising for tailoring standards to relevant context, but could spiral into a complex landscape of standards.

Stakeholders were positive about the concept of variable thresholds for minimum standards. Being able to tailor minimum standards to contextual factors for each technology was seen to be a good way to ensure that solutions are accredited based on their specific objectives, effectiveness, deployment context and not ruled out based on standards written with a different type of technology in mind.

Stakeholders also posited that variable thresholds may also make room for accrediting technologies with consideration of how they should be used, whether that is in specific circumstances or in combination with other approaches.

> **"We're almost giving more of a case study. So not saying accuracy of X, we're saying, accuracy of X for systems designed to do this thing, using this database, and in the context of a public post. You could start to build the metric and have lots of different elements of it. "**

Stakeholders commented on the contextual factors that may warrant variation in minimum standards:

- **Level of risk or harm:** Stakeholders felt that variable thresholds may allow the accreditation process to account for differing risk levels. For example, if a technology is designed to target types of content that poses an urgent risk of harm, it may be acceptable to have lower thresholds.

- **Harm type:** Stakeholders were in general agreement that technologies seeking accreditation should be focused on specific harm types, and developed to be effective and accurate in detection of that content. Therefore, they felt that there could be a case to vary minimum standards between the harm types in the same way, as the requirements and potential implications of targeting each harm will vary.

- **Public vs private content**: Stakeholders generally felt that thresholds on accuracy for technologies that access content communicated privately should be higher than those that only focus on content communicated publicly. They argued that automated tools deployed in the private context have a higher risk for causing harm through false positives and exposure of private content, and that variable thresholds could account for this key risk.

- **Severity of action after content flagged:** Viewing the application of an accredited technology as just a part of the overall process that a regulated service would need to follow, stakeholders felt that the acceptable accuracy of a technology would vary depending on the processes followed after potentially illegal content was detected. Some suggested that acceptable accuracy thresholds may be lower for some onward actions, for example referral to internal teams for further review. However, if the follow up action is more intrusive, such as reporting to law enforcement or taking action on an account, the threshold should be higher.

"I would argue that the minimum standard of accuracy that you want if an automated report was being made to law enforcement might be different from the minimum standard of accuracy that would be required to, for example, block a piece of content and pop up a warning."

- **Known vs unknown content**: Stakeholders discussed the challenges of detecting previously unknown CSEA content compared to content already reported, verified, and added to databases. Some suggested that technologies that are designed to identify unknown content could be subject to more lenient accuracy thresholds than those identifying already known content, to enable innovation of new solutions.

- **False negatives vs false positives**: Stakeholders discussed the different consequences of false negatives and false positives, and some felt that the accuracy thresholds for each of these should be different. Those who felt this way explained that while the implications of false negatives do mean that some harmful content is missed, a technology with a reasonable accuracy rate would still, overall, reduce the harmful content on the regulated service. On the other hand, the implications of a false positive can be very severe, potentially causing harm to individuals via exposing sensitive private content, prompting unfair sanctions by the service, or more significant legal and social consequences if a user is incorrectly accused of holding illegal content. Therefore, some stakeholders felt that accuracy thresholds for false positives must be much higher to mitigate these risks.

6. Operationalising the accreditation process

Chapter summary

Stakeholders felt that up-to-date, withheld datasets may be crucial to ensure robust accuracy testing as part of accreditation. However, they noted many limitations and challenges in creating and maintaining such a dataset, particularly where it needs to contain known and unknown CSEA or terrorism content.

There was suggestion of staggering the implementation of the accreditation process, to allow learning, and encourage early applications. For example, accreditation could begin for more established technologies that assess target content communicated publicly.

While stakeholders understood that the OSA requires a technology to be accredited to be eligible for a CSEA/terrorism tech notice, some suggested that further assessment might be needed to consider the specific deployment context ahead of a tech notice being issued. They thought this would be important to apply proper consideration to whether a technology is suitable for the service, harm type, and offender behaviour being targeted.

It was important to all stakeholders that any organisations involved in accreditation be trustworthy. There were mixed views on what types of organisations would be trustworthy to all parties, but there was some consensus that private companies may have commercial conflicts of interest. There was support for Ofcom conducting accreditation, possibly with the support of an independent third party.

Some stakeholders noted that accreditation timelines must consider the financial and resource burden on technology developers. They suggested that there should be a clear roadmap to accreditation, and clear expectations as to when and why the technology may be subject to re-accreditation. Stakeholders generally agreed that re-accreditation would be important to ensure that technologies continue to be sufficiently accurate.

Up-to-date, withheld datasets seen as crucial for robust accuracy testing as part of accreditation.

A theme which ran across the workshop was the importance of the having the right datasets to assess the accuracy of a particular technology against any metric-based minimum standards. During discussions about *how* the accreditation process should work, stakeholders returned to the importance of good quality data that provides confidence: both in terms of privacy and utility.

Stakeholders generally recognised that adequate data sets are key to developing and accrediting technology that is effective and accurate. However, some suggested that the limited availability of suitable datasets for training and testing technology would be a challenge to robust accreditation, and that adequate datasets might not exist at this time.

> **"Law enforcement has some datasets, companies have some datasets, clearing houses have some datasets, but none of the researchers or developers really have access to those types of datasets.… You would need to set up some type of trusted research environment where people from companies like you mentioned before who are vetted can do these types of training and evaluation. "**

Some stakeholders suggested that datasets should be regularly updated to reflect reality and must contain unknown content as well as known content to truly test effectiveness in detecting previously unknown content. They also raised practical issues associated with maintaining such databases; if the testing of new technologies that claim to detect unknown content successfully does so, the content would then need to be reported (a legal obligation for CSEA content in particular), making it then 'known'. This would over time reduce the proportion of unknown content represented in the data, requiring that the dataset be replenished.

Stakeholders felt that such a dataset would have to be securely held by Ofcom or an entity appointed by Ofcom, with no other parties able to access the dataset. This was seen as key to protecting the integrity of accuracy tests done using the dataset. They suggested that this approach would improve confidence from regulated services, users, and civil society groups with regards to the privacy and safety implications of accredited technology.

Some suggestion of staggering the implementation of the accreditation process to allow learning, and encourage early applications.

Some stakeholders suggested that Ofcom could start with accreditation of established technologies such as cryptographic hashing products on content communicated publicly, or non-encrypted private content. They thought that this would allow the accreditation process to be operationalised on technologies that are already widely adopted first and apply the learnings from the process to more challenging contexts/technologies later on. Stakeholders felt a benefit of this staggered approach would be that learnings from the earlier applications could then be applied for the more challenging contexts later on.

Stakeholders raised the importance of making accreditation attractive for prospective applicants to produce tools and encourage innovation, citing that process burdens could influence commercial decisions about whether to seek accreditation. With this in mind, stakeholders said that there must be a clear roadmap to outline to prospective applicants *how* to become accredited, to help them mitigate the risk of investing in development and ultimately being unsuccessful in seeking accreditation.

While a technology must be accredited to be eligible for a tech notice, further assessment might be needed to consider the specific deployment context ahead of a tech notice being issued.

Some stakeholders described a catch-22 situation: Ofcom being unable to issue tech notices for technology that is not accredited, meaning that technology has to be accredited without full understanding of the context in which it might ultimately be required.

Others flagged the difference between how standards are set and how they are applied, suggesting that it may be that a lower standard is required to become accredited, but secondary evaluation or assessment may be needed considering application of that accredited tech in a specific circumstance to judge how accurate and appropriate it is for that harm/service.

There was also a suggestion that the accreditation process itself could mirror variable minimum standards, in that there are different processes for different contexts. For example, a technology designed only to identify target content communicated publicly could be subject to a lighter touch accreditation process, while technology that identifies content communicated privately could be subject to a much more thorough accreditation process.

The organisation(s) involved in accreditation must be trustworthy.

Stakeholders reflected on the fact that the OSA provides flexibility for either Ofcom, or another person appointed by Ofcom, to accredit technology against the minimum standards of accuracy. There was no consensus around which organisation(s) should be responsible for accreditation, however stakeholders agreed around the qualities needed in such organisation:

- A trusted independent partner
- A legitimate voice
- A technical skill set with adequate resource
- Access to appropriate information, for example governance structures or training datasets where relevant.

> **"I think it's trying to work out who everyone thinks is a legitimate source and if there isn't, which is likely the case, then it goes back to Ofcom".**

Stakeholders cautioned using private companies to conduct the accreditation process given the potential for commercial conflicts of interest and risks around handling highly sensitive information. Stakeholders were often in favour of Ofcom taking on the job of accrediting technology given their independence from Government and lack of commercial conflict.

> **"Personally, I think Ofcom doing that would be probably the safest option, they're a regulator, you trust them. But then on the other hand, as you already mentioned, you know, the resource burden, so this is something that would need to be considered. Like, is it actually feasible for Ofcom to carry that out?"**

Across the room, stakeholders flagged the importance of including a wide range of voices in the accreditation process in the same way the workshop had done. One suggestion was to do so through the role of an advisory group, with representation from academia, child safety groups and counter terrorist groups, although some cautioned that this might introduce bias.

> **"I do think that independent validation from difference sources is right … I also think consulting with experts, and it's great that you've had academics here today who have really dedicated their lives to thinking about these problems".**

Timelines must consider burden on technology developers, but accreditation must be periodically reviewed.

Stakeholders talked about the importance of re-accreditation of technology, given the ever-changing landscape. One suggestion was for a periodic re-accreditation of technologies, with the option to carry out spontaneous re-assessments if there are indications of issues with a particular accredited technology.

Stakeholders pointed to the risk of re-accreditation during the service of a tech notice. For example, if a tech notice was issued for a defined period of time but the accreditation of the required technology expired before it elapsed, stakeholders suggested it could create a problem for the company subject to the notice.

> **“Say you issued a tech notice that lasted a year but you were using an accredited technology whose accreditation ran out 2 months into that year, how do you align the time of the re-accreditation?”**

One stakeholder suggested the approach to re-accreditation could involve a three-stage process. The first stage would involve a baseline accreditation for services or companies, followed by the development and execution of an action plan by the applicant. Finally, the applicant would go through a re-accreditation process to ensure ongoing adherence to minimum standards throughout the term of the notice.

Stakeholders recognised that the accreditation process could be time-consuming and resource intensive, and cautioned about the risk of it disincentivising prospective applicants from seeking accreditation.

> **“The process, I would assume, would take a long from the beginning, when you start your accreditation, until you finalise it. And then, let's say it takes a year or something. By the time they're done, if there's another re-accreditation and another year long process…You have to have extra resources for that, you would need to potentially recruit people to help you to comply with that and so on.”**

Relevant standards and accreditations

The following standards were met by Ipsos in carrying out this project:

ISO 20252

This is the international specific standard for market, opinion and social research, including insights and data analytics.

Market Research Society (MRS) Company Partnership

By being an MRS Company Partner, Ipsos UK endorse and support the core MRS brand values of professionalism, research excellence and business effectiveness, and commit to comply with the MRS Code of Conduct throughout the organisation & we were the first company to sign our organisation up to the requirements & self-regulation of the MRS Code; more than 350 companies have followed our lead.

ISO 9001

International general company standard with a focus on continual improvement through quality management systems.

ISO 27001

International standard for information security designed to ensure the selection of adequate and proportionate security controls.

The UK General Data Protection Regulation (UK GDPR) and the UK Data Protection Act 2018 (DPA)

Ipsos UK is required to comply with the UK General Data Protection Regulation and the UK Data Protection Act; it covers the processing of personal data and the protection of privacy.

HMG Cyber Essentials

A government backed and key deliverable of the UK's National Cyber Security Programme. Ipsos UK was assessment validated for certification in 2016. Cyber Essentials defines a set of controls which, when properly implemented, provide organisations with basic protection from the most prevalent forms of threat coming from the internet.

Fair Data

Ipsos UK is signed up as a 'Fair Data' Company by agreeing to adhere to twelve core principles. The principles support and complement other standards such as ISOs, and the requirements of Data Protection legislation. .

For more information

3 Thomas More Square
London
E1W 1YW

t: +44 (0)20 3059 5000

www.ipsos.com/en-uk
http://twitter.com/IpsosUK

About Ipsos Public Affairs

Ipsos Public Affairs works closely with national governments, local public services and the not-for-profit sector. Its c.200 research staff focus on public service and policy issues. Each has expertise in a particular part of the public sector, ensuring we have a detailed understanding of specific sectors and policy challenges. Combined with our methods and communications expertise, this helps ensure that our research makes a difference for decision makers and communities.