Your response

Note On This Response

This response represents a view of the membership of OSTIA determined through consultation, and issues raised here cut across the Online Safety Tech community.

The majority of the material in this response is sourced from OSTIA members, many of whom may also have made individual responses making similar comments. While there may be duplication from individual responses, we believe it is helpful to distinguish the views of individual members from this sectoral response.

Volume 2: The causes and impacts of online harm

Ofcom's Register of Risks

Question 1:	
i)	Do you have any comments on Ofcom's assessment of the causes and impacts of online harms?
Response:	
ii)	Do you think we have missed anything important in our analysis? Please provide evidence to support your answer.

Response:

Sufficiently Well Defined Harms

In our work with policing, NGOs and policymakers we often find that "CSAM" is treated as a harm in its own right. When designing systems with "safety by design" in mind, or considering moderation and mitigations, we believe that this is unhelpful. We believe that under the category of illegal CSAM there are a number of distinct harms, for which mitigations whether proactive or reactive must be considered separately.

OfCom's analysis (Volume 2, Page 63) hints at these of the distinct harms in its introduction and later in sections 6C.108 to 6C.123 , recognising the trauma to victims and survivors, revictimization, and unintentional viewing. I would suggest that these should be given more consideration and more detail provided. While we are NOT the experts on how this should be accomplished, we will share some of distinct harms that we use in our internal analysis when considering mitigations. I recognise that the document mentions many of these, but in passing rather than in a structured way.

- **Harm to Victims and Survivors**

- Knowledge that the abuse is being viewed or seen by others is in itself traumatising.
- Survivors may be further confronted with the abuse and experience new trauma when they become aware images or videos are continuing to circulate, for example when they are cited in court cases.
- Offenders viewing material will sometimes choose to contact survivors who they have seen being abused. Some find gratification from letting people know they have seen it. Others will attempt to exploit what they perceive as vulnerability in the survivor to carry out further abuse of the survivor, even if they are an adult. Survivors report being recontacted in this way is highly traumatic.
- As a result of all of the above survivors often live in fear of being recognised in public or online by offenders who have seen the abuse through CSAM, with serious implications for wellbeing and ability to participate fully in society.

- **Harm to Unintentional Viewers**
  - People who are unintentionally exposed to CSAM may be traumatised by that experience.
  - Some people who are unintentionally exposed to CSAM may forward that material to others out of indignation, outrage or seeking help. This may traumatise others, and may contribute to harm to victims and survivors (see above) by making the CSAM accessible to more people.
  - Some people who are unintentionally exposed to CSAM may have a latent potential for an interest in CSAM or the abuse of children which is triggered by an initial unintended exposure and causes them to go on to seek out more CSAM content contributing to harm to victims and survivors (see above).
  - Some people who are unintentionally exposed to CSAM may react to the trauma they experience by developing an addiction to CSAM content contributing to harm to victims and survivors (see above).

- **Harm Escalation** (CSAM Radicalisation Pathway)
  - For those who have an interest in CSAM, wide availability of CSAM can contribute to normalising the idea that it's OK. "It's on *MainstreamPlatform* so it can't be that wrong". This may lead them to continue and escalate their offending.
  - Communities formed around CSAM content online tend to encourage those with an interest in CSAM to normalise that interest and escalate their offending.
  - Communities formed around CSAM content online may encourage participants to commit escalating offences including accessing more material, accessing material of a more extreme or severe natures, and committing offences against children either online in person (including CSAM related and grooming, livestreaming, contact abuse).
  - Some communities, often on the dark web, require prospective members to provide newly generated CSAM as the "price of entry" and to prove they are not working against the members of that community (e.g. for Police or NGO). This can encourage or coerce direct offences against children that might not otherwise have taken place.
  - The collective impact of these harms is expanding the overall volume of offenders and offending.

- **Potential Harm to Children**
    - A number of studies show that many of those accessing CSAM online have thoughts about committing offences against children online or in person. Some of those people will translate those thoughts into action and create new victims and survivors, or inflict new abuse on victims.
    - Where there is a "market" for CSAM there will always be those who seek to serve it whether for recognition or profit. This results in harm to children as new CSAM is created.

Understanding these different harms is fundamental to designing safe systems and assessing the adequacy of responses and mitigations. This is because mitigations typically work only for a subset of harms, so multiple mitigations are needed to be effective against the full range of harms. Some of those mitigations have higher cost and more tradeoffs (e.g. with privacy or user experience) than others.

We often see discussions miss opportunities for positive interventions when people are talking at cross purposes about the harm being targeted. For example, identifying known CSAM is highly effective in reducing Harm to Survivors of that abuse, but has no immediate impact for survivors where the related CSAM is unknown – although once that material becomes the measure can become effective after a lag.

We propose that the code is updated to reflect this type of approach, ideally with input from stakeholders with deep understanding of this area and bringing in survivor perspectives. Much of the data to support this approach is already referenced on subsequent pages, and while a number of areas are explicitly broken down (e.g. types of SGII) we think further work is required for CSAM more generally, and for terrorism related harms.

We believe that fine-grained consideration of harms is vital to effective governance, risk assessment and mitigation.

iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response: No	

Question 2:	
i)	Do you have any views about our interpretation of the links between risk factors and different kinds of illegal harm? Please provide evidence to support your answer.
Response: We refer to our answer to the previous question. We believe analysis of risk factors gives insufficient consideration to the different harm types arising from those risk factors.	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response: No	

Volume 3: How should services assess the risk of online harms?

Governance and accountability

Question 3:	
i)	Do you agree with our proposals in relation to governance and accountability measures in the illegal content Codes of Practice?
Response:	
ii)	Do you think we have missed anything important in our analysis? Please provide evidence to support your answer.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 4:	
i)	Do you agree with the types of services that we propose the governance and accountability measures should apply to?
Response:	
ii)	Please explain your answer.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 5:	
i)	Are you aware of any additional evidence of the efficacy, costs and risks associated with a potential future measure to requiring services to have measures to mitigate and manage illegal content risks audited by an independent third-party?
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 6:	
i)	Are you aware of any additional evidence of the efficacy, costs and risks associated with a potential future measure to tie remuneration for senior managers to positive online safety outcomes?
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Service's risk assessment

Question 7:	
i)	Do you agree with our proposals?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Specifically, we would also appreciate evidence from regulated services on the following:

Question 8:	
i)	Do you think the four-step risk assessment process and the Risk Profiles are useful models to help services navigate and comply with their wider obligations under the Act?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 9:	
i)	Are the Risk Profiles sufficiently clear?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Do you think the information provided on risk factors will help you understand the risks on your service?
Response:	
iv)	Please provide the underlying arguments and evidence that support your views.
Response:	
v)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Record keeping and review guidance

Question 10:	
i)	Do you have any comments on our draft record keeping and review guidance?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 11:	
i)	Do you agree with our proposal not to exercise our power to exempt specified descriptions of services from the record keeping and review duty for the moment?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Volume 4: What should services do to mitigate the risk of online harms

Our approach to the Illegal content Codes of Practice

Question 12:	
i)	Do you have any comments on our overarching approach to developing our illegal content Codes of Practice?
Response:	

We wholeheartedly agree with and support the OSA Network statement on the Illegal Harms Consultation, which can be found here: https://www.onlinesafetyact.net/analysis/osa-network-statement-on-illegal-harms-consultation/ . They express a set of concerns we fully agree with, and more eloquently that we could.

We would like to expand on the point about focus on best practice. The Government stated throughout the process of passing the Online Safety Act that most platforms were not doing enough. This certainly came from the premise that best practice as accepted by industry today falls short. By focusing on achieving best practice on a slightly wider scale this consultation fails to deliver against many of the aspirations that drove the passing of the Act.

There is also a dangerous circular logic. There is little or no incentive in this consultation for industry to improve best practice (merely achieve it). As a result, best practice is likely to remain static, and future reviews of this guidance will not "raise the bar" as best practice has not changed. This is a missed opportunity to have the Act and Codes of Practice drive improvement.

ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response: No	

Question 13:	
i)	Do you agree that in general we should apply the most onerous measures in our Codes only to services which are large and/or medium or high risk?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 14:

i)	Do you agree with our definition of large services?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 15:	
i)	Do you agree with our definition of multi-risk services?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 16:	
i)	Do you have any comments on the draft Codes of Practice themselves?
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 17:	
i)	Do you have any comments on the costs assumptions set out in Annex 14, which we used for calculating the costs of various measures?
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Content moderation (User to User)

Question 18:	
i)	Do you agree with our proposals?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Content moderation (Search)

Question 19:	
i)	Do you agree with our proposals?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Automated content moderation (User to User)

Question 20:	
i)	Do you agree with our proposals?
Response: No	
ii)	Please provide the underlying arguments and evidence that support your views.

Response:

Regulated Services

Dame Melanie Dawes stated in the Public Accounts Committee that the initial objective in implementing the Online Safety Act is "is to take what the industry is already doing, put the evidence behind it, and then get everybody doing it, in order to raise the bar and raise the standard".

This section of the code appears to take the level of intervention or mitigation required from "is to take what the industry is already doing, put the evidence behind it, and then get everybody doing it". This will undoubtedly lead to some improvements, but as most measures, even baseline capabilities such as hash matching of known CSAM, are only required for large and high risk platforms it fails to deliver on the intent to "get everybody doing it". Ofcom's analysis of "The causes and impacts of online harm" recognises evidence that offenders often use smaller, less well moderated, U2U services to host and promote content, often linking it (directly or indirectly) from larger platforms, and recognises this as a risk factor. The proportionality of regulation is important, but in other areas of safety we do not allow companies to avoid baseline requirements simply on the basis of size. If you wish to market an electrical appliance, vehicle, food or medicine to the public there are minimum standards for companies at all sizes. Given the volume of evidence Ofcom has presented across all harms on the role of smaller platforms, we believe the current proposals are far too weak.

Requirement for Hashing

With respect specifically to measures related to hashing, the section 14 on Automated Moderation acknowledges this:

> "In principle, we provisionally consider that, even where they are very small, it could be justified to recommend that services which are high risk to deploy these technologies. However, we are proposing to set user-number thresholds below which services would not be in scope of the measure. This is because to implement hash matching and URL detection services will need access to third party databases with records of known CSAM images and lists of URLs associated with CSAM. There are only a limited number of providers of these databases, and they only have capacity to serve a finite number of clients. Setting the user-number thresholds we have proposed should ensure that the database providers have capacity to serve all services in scope of the measure. Should the capacity of database providers expand over time, we will look to review whether the proposed threshold remains appropriate."

To expand the capacity of database providers is likely to require investment. It is unclear whether the statement "...we will look to review…" is adequate to enable that investment. Some of the investment required will be within the database providers, which are typically NGOs already struggling with funding in a difficult economic environment. Other investment may be required in companies in the private sector which typically provide the engines of innovation creating new products and services to enable expansion of capacity – whether within regulated services or in the wider commercial community. Such investment is typically made based on a clear expectation of outcome, whether through social impact, financial returns or regulatory compliance, either by organisations themselves or by their funders. As structured just now, there is a danger we remain trapped in a vicious cycle:

- Investment in scaling availability of hashing has not been made as the demand does not justify the required investment.
- Ofcom has indicated it cannot regulate without that scaled availability being in place, with no indication of when such a review might take place or what would be required.
- Were such a review to take place there is no surety Ofcom will recommend a change, and even if a change were proposed it could potentially be a small change bringing a handful of additional platforms into scope.

This does not create a driver for investment to break the cycle.

Ofcom should make a clear statement, such as "We will review the capacity of database providers annually and hope to be able to bring Xxx additional platforms in scope for this requirement within 6 months of capacity becoming available". This would provide a much clearer basis for the investment needed for future expansion.

Any clarity Ofcom can provide would help, including

- Which services (size, profile) Ofcom would have sought to place this requirement on if capacity were not an issue
- Under what circumstances and on what timescale a review might take place
- What factors the review might consider
- What Ofcom would be seeking to achieve with the review

Clarity could create the incentive needed for investment in change.

Promoting Online Safety Innovation

Dame Melanie Dawes also referred commitments to "raise the bar" and "raise the standard". A huge volume of discussion in and around the parliamentary process for what is now the Online Safety Act centred on how new technology could improve online safety, and empowering a regulator to require the use of such technology where appropriate.

We believe that technology is a necessary component of improving online safety because of the need to operate at scale. Technology can also operate in ways the protect privacy by avoiding unnecessary moderator viewing of private content. We know that deploying people as moderators has a human cost as well as an economic one. Moderators frequently report experiencing poor mental health and trauma from the content they are required to review, and there have also been reports of suicide, and of moderators becoming addicted to exactly the sort of toxic content they are paid to remove. Technology that can minimise and support human intervention is crucial.

For these reasons the UK government, through DSIT and the Home Office, has continued to express a strong desire for innovation in online safety technology. This is re-affirmed in the recent MoU with the Australian government. In doing so it has identified areas where it would like to see innovation and promoted Online Safety as an area for investment for innovation at all stages from fundamental research in Universities through to commercial development at later Technology Readiness Levels.

The technologies mentioned in the section on Automated Content Moderation of this consultation are not at the cutting edge of innovation:

- Hash matching dates back to 1979 and has been in use for CSAM since at least the 1990s
- Perceptual hashing dates back to 1980 and PhotoDNA for CSAM to 2009
- Keyword matching is almost as old as computing – it would have been familiar in Bletchley Park in WWII
- URL matching is almost as old as the internet, dating back to the 1990s

There is a real danger that this signals to regulated services that they can rely on old technologies, and that there is no need for them to invest in newer technologies either through internal development or by buying in. This also fails to signal to investors in innovation elsewhere, from research councils and Innovate UK to private sector Angel and Venture investors, that there is any incentive to create new technology.

We have heard from Ofcom on a number of occasions that they will seek to continuously review these codes and "raise the bar". However, there is little clarity here or elsewhere on how this will happen.

We believe that a high functioning innovation ecosystem for Online Safety would:

- Have a clear understanding of where the regulator would like to be able to act, and how and when it would be able to do so.
- Feel confident to invest in the development of new online safety technologies to the point where there is clear evidence for their efficacy in addressing harm
- Have a reasonable expectation that the regulator would encourage or mandate the use of technologies with a sufficient evidence base
- Have a reasonable expectation that the actions of the regulator would create conditions where the technology would be able to deliver the desired outcomes in online safety and provide return on investment (whether measured in social outcomes or financial ones).

Currently none of these conditions is true.

We would therefore encourage Ofcom to indicate now:

- How and when codes will be revised
- Priority areas where Ofcom would be keen to recommend or require use of automated moderation technology were it available, and what evidence would be required

This would play a significant role in enabling investment in technology development aligned with Ofcom's goals, creating a virtuous cycle of innovation and mitigating the damage done to the online safety technology ecosystem by these initial draft codes.

While slightly tangential to this consultation, it is also worth noting that outside of regulated services, the availability of data to determine technical approaches, train, and/or test innovation is often unavailable. It is very hard to build tools to detect harm if there is only anecdotal data about how that harm takes place. We encourage Ofcom to consider whether it can contribute to bringing together innovators and harms insight, training and/or test data to help build a high functioning ecosystem.

References to Specific Proposals

14.26 False Positives

When considering the performance of detection technologies "false positive rate" is only one relevant component.

- **False Positive Rate** tells us how many false positives will be generated for a given volume of content
- **Nature of False Positives** tells us the characteristics of these false positives
- **Consequence** allows us to explore what the impact of these false positives is on the users rights.

Assessment of false positive rate without consideration for the nature of false positives and what the action following detection (and consequence thereof) is meaningless, and the consultation documents should reflect this.

14.53

"Further, we are aware of recent research that has indicated perceptual hashing algorithms could be repurposed to add hidden secondary capabilities.[197]"

We believe this statement to be inconsistent with the terminology in this document and therefore incorrect.

The use of the term "perceptual hashing" up to this point in the document appears to describe technologies such as PhotoDNA which match a specific hash against a database, usually using some form of Euclidean distance. These solutions rely on the integrity of the database, and the measures for ensuring the integrity of the database are described elsewhere in this consultation. There are multiple ways of verifying end-to-end integrity of the system as the original database entries can be reviewed by humans, and map one-to-one onto hashes through a deterministic mathematic process. Either exact matching or a well know heuristic (typically Euclidian distance) is used to determine matches.

The paper referenced at 197 does not describe perceptual hashing in this sense. Instead it describes a system where a deep learning model is fed with CSAM images non CSAM images to "train" it. The output of this training process is a "model", often built on top of a base model.

This model cannot usually be mapped back onto the training data in any human understandable way, and the authors of the paper demonstrate that a "dual purpose" model can be built which is effectively indistinguishable from a single purpose one. The lack of explainability and transparency is a risk across many AI and Machine Learning technologies. We believe this is, from a technology and impact perspective, very different to "perceptual hashing.

We believe the statement should more correctly read:

"Further, we are aware of recent research that has indicated deep learning algorithms could be repurposed to add hidden secondary capabilities.[197]"

This might be a justification for suggesting caution in the use of such models, but is irrelevant to the performance of perceptual hashing as described elsewhere in this document.

We are making the assumption that it is not Ofcoms intention to include deep learning models within the term "perceptual hashing". If it is Ofcom's intention that deep learning approaches should be seen as a form of perceptual hashing then many of the other statements about performance require significant revision to reflect characteristics of deep learning models including lack of explainability and traceability, and risk of unintended bias.

14.54 Biases

This is an excellent description of the likely biases in hashing or perceptual hashing approaches. We suggest it would be useful to consider the consequences of these biases on different harms.

Public discussion of bias often focuses on disadvantage and exclusion which relates to the creation of new harms or amplify existing biases (and prejudices). Typically we are talking about the introduction of a measure that leaves some group in society worse off than they were under the previously existing measures.

The introduction of perceptual hashing as Ofcom proposes does not have a primary effect of making any group in society worse off than if detection were not introduced. There could be a second order effect that if offenders understood that CSAM containing some groups were less detectable they would specifically target that group to evade detection increasing inequality. While still undesirable and something that should be used to drive continuous improvement, this is not the same as the type of bias which excludes groups from participation or targets them unfairly.

We assume from context that Ofcom has taken this into consideration, but we believe that this subject should be covered explicitly, probably at this point 14.54 so that at the same time as acknowledging potential bias, Ofcom explains its reasons for believing this level of bias is not itself a barrier to deployment.

We also suggest that explicitly stating databases must avoid systematic bias within their control would be helpful. For example by adding a statement in A15.23 that databases should determine addition of content solely based on whether or not it is CSAM, and ensure minimisation of bias in processes making that determination. If some database systematically refused to include CSAM relating to a particular gender, sexuality or ethnic group it should be clear that was not acceptable to use that database for the purposes of complying with this regulation. For the avoidance of doubt we have no reason to believe any such bias exists in any database today.

14.109

There is an implicit assumption that a new risk assessment on the service deemed low risk would identify the presence of CSAM and thus increased risk of CSAM in future.

If there is no pro-active detection of CSAM and no pro-active moderation, how would the low risk service know that it had been used for CSAM?

This information could come into being if a user report had been made, but even in public services there is often the ability to create content in such a way that it is hard to discover even though Appendix 9s guidance would suggest that it has been "communicated publicly" for the purposes of the act. Such content can then be shared by a group of offenders, none of whom is likely to report it. Unless another user stumbles across it, no report will be made.

We therefore believe that the assertion that "new risk assessment would identify CSAM" is likely to be incorrect in most practical cases. In practice, small platforms could easily be oblivious that they were being used by offenders. We believe this is a compelling reason to increase the scope of application for hashing in automated moderation to a wider proportion of platforms, including smaller ones.

Mitigating Harms from SGII

Volume 2 of the consultation document recognises SGII (6C) recognises the increasing prevalence of SGII and the harm it causes, yet the recommendations for Automated Content Moderation do offer nothing to combat this type of harm except where imagery is already in hash lists (by which time huge harm has doubtless already occurred.

Preventing SGII should be a priority, and we believe that for large or high risk platforms there are measures that could be recommended by Ofcom.

- **Prevention of SGII**
    - Age Assurance technology is already recognised by Ofcom as an effective solution in the recommendations relating to pornography. If this same technology were applied on regulated services, it would be possible to reliably identify which users are children.
    - There are very few legitimate use case for children to either post content containing nudity. There are classifiers which can identify nudity to a very high degree of confidence.
    - Combining age assurance (knowing which users are children) with nudity detection (focused on those accounts) could play a hugely important role in preventing children posting content containing nudity including SGII
    - This could be used block upload with an option to appeal blocking with a moderator. In this mode the SGII is not seen by anyone protecting the child's privacy.
    - An alternative would be conventional referral to moderator for review, with the consequence that at least one person will view the SGII resulting in some intrusion to a childs privacy (albeit with good intentions)
    - We would hope that in either case the response be appropriate and ensure that children were supported and not criminalised for their actions, which are almost always misguided or coerced rather than criminal in intent.
    - Nudity detection is not 100% reliable (no technology is) but the consequences of a false positive are limited if there is moderator review or appeal available. By limiting application of the technology to accounts of children, false positives have no impact on the adult population.
- **Identification of Previously Unknown CSAM including SGII**

- New imagery including SGII will not be detected by hashing until it has been discovered through other means and added to databases.
- There are classifiers which are focused on detecting CSAM.
- The accuracy of these detectors is sufficient that they should play an important supporting role for large or high risk platforms in detecting CSAM including SGII
- For platforms that do not permit nudity accuracy is very likely to be high, as false positives tend to come from misinterpretation of age rather than activity.
- For platforms that do allow nudity it may be appropriate to set higher thresholds and/or use approach based on "strikes" (where moderator intervention occurs when a certain number of items are flagged), which could happen at the level of user accounts or groups.
- Information from detection could also be combined with other risk factors for individual users or groups. Platforms have publicly claimed to use factors including behaviour, behaviour based age estimation, and metadata including which users are communicating with each other to generate risk scores, and CSAM classifiers would clearly add valuable information to such tools.

Given the availability and level of demonstrated reliability and efficacy of these technologies both in online applications and in Law Enforcement, we believe these should at the very least be referred to as a route that large or high risk platforms should consider using in mitigating risk relating to SGII and previously unknown CSAM more generally, with a requirement to demonstrate effective alternative capabilities to detect these categories of content if these measures are not adopted.

iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response: No	

Question 21:	
i)	Do you have any comments on the draft guidance set out in Annex 9 regarding whether content is communicated 'publicly' or 'privately'?
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Do you have any relevant evidence on:

Question 22:	
i)	Accuracy of perceptual hash matching and the costs of applying CSAM hash matching to smaller services;
Response:	

ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 23:	
i)	Ability of services in scope of the CSAM hash matching measure to access hash databases/services, with respect to access criteria or requirements set by database and/or hash matching service providers;
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 24:	
i)	Costs of applying our CSAM URL detection measure to smaller services, and the effectiveness of fuzzy matching for CSAM URL detection;;
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 25:	
i)	Costs of applying our articles for use in frauds (standard keyword detection) measure, including for smaller services;
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 26:	
i)	An effective application of hash matching and/or URL detection for terrorism content, including how such measures could address concerns around 'context' and freedom of expression, and any information you have on the costs and efficacy of applying hash matching and URL detection for terrorism content to a range of services.
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Automated content moderation (Search)

Question 27:	
i)	Do you agree with our proposals?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

User reporting and complaints (U2U and search)

Question 28:	
i)	Do you agree with our proposals?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Terms of service and Publicly Available Statements

Question 29:	
i)	Do you agree with our proposals?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 30:	
i)	Do you have any evidence, in particular on the use of prompts, to guide further work in this area?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Default settings and user support for child users (U2U)

Question 31:	
i)	Do you agree with our proposals?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 32:	
i)	Are there functionalities outside of the ones listed in our proposals, that should explicitly inform users around changing default settings?
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)

Response:

Question 33:	
i)	Are there other points within the user journey where under 18s should be informed of the risk of illegal content?
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Recommender system testing (U2U)

Question 34:	
i)	Do you agree with our proposals?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 35:	
i)	What evaluation methods might be suitable for smaller services that do not have the capacity to perform on-platform testing?
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

We are aware of design features and parameters that can be used in recommender system to minimise the distribution of illegal content, e.g. ensuring content/network balance and low/neutral weightings on content labelled as sensitive.

Question 36:	
i)	Are you aware of any other design parameters and choices that are proven to improve user safety?
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Enhanced user control (U2U)

Question 37:	
i)	Do you agree with our proposals?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 38:	
i)	Do you think the first two proposed measures should include requirements for how these controls are made known to users?
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 39:	
i)	Do you think there are situations where the labelling of accounts through voluntary verification schemes has particular value or risks?
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

User access to services (U2U)

Question 40:	
i)	Do you agree with our proposals?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Do you have any supporting information and evidence to inform any recommendations we may make on blocking sharers of CSAM content? Specifically:

Question 41:	
i)	What are the options available to block and prevent a user from returning to a service (e.g. blocking by username, email or IP address, or a combination of factors)?
Response:	
ii)	What are the advantages and disadvantages of the different options, including any potential impact on other users?
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 42:	
i)	How long should a user be blocked for sharing known CSAM, and should the period vary depending on the nature of the offence committed?
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

There is a risk that lawful content is erroneously classified as CSAM by automated systems, which may impact on the rights of law-abiding users.

Question 43:	
i)	What steps can services take to manage this risk? For example, are there alternative options to immediate blocking (such as a strikes system) that might help mitigate some of the risks and impacts on user rights?
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Service design and user support (Search)

Question 44:	
i)	Do you agree with our proposals?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Cumulative Assessment

Question 45:	
i)	Do you agree that the overall burden of our measures on low risk small and micro businesses is proportionate?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 46:	
i)	Do you agree that the overall burden is proportionate for those small and micro businesses that find they have significant risks of illegal content and for whom we propose to recommend more measures?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 47:	
i)	We are applying more measures to large services. Do you agree that the overall burden on large services proportionate?
Response:	

ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Statutory Tests

Question 48:	
i)	Do you agree that Ofcom's proposed recommendations for the Codes are appropriate in the light of the matters to which Ofcom must have regard?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Volume 5: How to judge whether content is illegal or not?

The Illegal Content Judgements Guidance (ICJG)

Question 49:	
i)	Do you agree with our proposals, including the detail of the drafting?
Response:	
ii)	What are the underlying arguments and evidence that inform your view?
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 50:	
i)	Do you consider the guidance to be sufficiently accessible, particularly for services with limited access to legal expertise?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 51:	
i)	What do you think of our assessment of what information is reasonably available and relevant to illegal content judgements?
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Volume 6: Information gathering and enforcement powers, and approach to supervision.

Information powers

Question 52:	
i)	Do you have any comments on our proposed approach to information gathering powers under the Online Safety Act?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Enforcement powers

Question 53:	
i)	Do you have any comments on our draft Online Safety Enforcement Guidance?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Annex 13: Impact Assessments

Question 54:	
i)	Do you agree that our proposals as set out in Chapter 16 (reporting and complaints), and Chapter 10 and Annex 6 (record keeping) are likely to have positive, or more positive impacts on opportunities to use Welsh and treating Welsh no less favourably than English?
Response:	
ii)	If you disagree, please explain why, including how you consider these proposals could be revised to have positive effects or more positive effects, or no adverse effects or fewer adverse effects on opportunities to use Welsh and treating Welsh no less favourably than English.
Response:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	