Your 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.

For example, in later volumes there is a recommendation regarding the use of hashes to detect known CSAM. Hashes can be used to trigger a number of different responses including **block**ing content, **warn**ing users about the consequences of their behaviour and/or **signpost**ing sources of help to get out of a cycle of offending, and generating **report**s to law enforcement. Even within the use of hashing, these different responses to detection are effective in different ways and to different extents for different harms.

For example, some of the positive effects of these approaches might be viewed as follows:

Impact	Block	Warn / Signpost	Report
Harm to Survivors	**+++** Reduced viewing of CSAM	**++** Reduced risk of exposure	**+** Long term reduction in offending
Harm to Unintended Viewers	**+++** Reduced risk of exposure	**++** Reduced risk of exposure	**+** Long term reduction in offending
Harm Escalation	**+++** Disrupts radicalisation	**++** Disrupts radicalisation	**+++** if offenders are caught
Potential Harm to Children	**+** Long term reduction in offending	**+** Long term reduction in offending	**+++** if offenders are caught

Similar evaluation could be considered on, for example, AI for detecting previously unknown material, interventions at community level rather than content/user level etc.

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 it becomes known it is effective after a lag.

We emphasise that we are not proposing this as a complete analysis nor do we claim to be expert in this field. There are others in academia and NGOs in particular who would have considerable expertise in this area, and far more able to bring evidence of offender and offence archetypes and to bring survivor centric approaches. We simply wish to illustrate the point that CSAM is not a single monolithic harm (even when limited to known CSAM) and that it should not be treated as such in risk assessment or assessment of the effectiveness/appropriateness of mitigation.

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) I think further work is required for CSAM more generally.

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.

We will use a grossly oversimplified example to illustrate the importance of the nature of false positives.

Most systems permit tuning of the false positive rate. In cryptographic hashing this is effectively controlled by the hash length. In perceptual hashing, AI or machine learning based systems a threshold of some sort is applied.

Imagine that we have two systems for CSAM detection, one cryptographic and one based on AI, which have been tuned to have the same false positive rate of 1%. A moderator reviews content flagged by the detector.

Imagine a user accessing a variety of services and participating in a variety of communities and conversations sends 100 images in a week. 90% of these images are mundane memes, pictures of their friends, their cat, and meals. 10% if the images they send are sensitive and highly personal intimate images exchanged with a partner and pictures of their baby at bathtime showing nudity. All of this activity is legal, and the user has a high expectation of privacy especially in the latter category of sensitive images.

The cryptographic hashing technology could be expected to generate one false positive for review by a human moderator in a week. There is a 90% chance that the image is mundane and although our user may be mildly annoyed by a moderator seeing their image it is not a matter of great concern to them. The overall probability of a moderator seeing a sensitive image in a week is around 0.1%.

The AI technology is using cues such as nudity, sexual acts or poses, and age to infer the presence of CSAM. It is far more likely to generate a false positive from an intimate image exchanged with a partner or a picture of a baby at bathtime than it is from a political meme. The overall probability of a moderator seeing a sensitive image is closer to 10x higher at 1% (most if not all false positives will be sensitive). Our user is likely to be outraged that this technology has not only sent his content to a moderator but it has focused in on his most private content.

Clearly the scenario is contrived to illustrate the point and real world number for the two systems would likely be very different (cryptographic systems are typically tuned for better than 0.000001% false positives, AI struggles to achieve less than 5% in these applications), but the problem is very real. False positive rate is meaningless unless we also consider the nature of the false positives generated. For Cryptographic hashing these are completely random. For perceptual hashing with usual thresholds slightly less so, for AI based technologies detection will tend strongly towards sensitive content.

The consequence must also be considered. In the contrived example above the outrage of the user is not because of a false positive in the detection system – it is from another person seeing their private content. Had the consequence of the false positive been that they were blocked from sending an image they may be irritated but is unlikely to feel the same anger.

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.

14.24 – Hashing and Privacy

Regulated services will need to consider compliance with Data Protection regulations in addition to those relating to Online Safety. Both Cryptographic Hashes and Perceptual Hashes can be personally identifying information under GDPR.

There is reference in A15 to both PhotoDNA and GoogleCSAI. Both of these API based services currently require transmission of either the original images or perceptual hashes to a third party for processing. There is a risk to privacy from an actor who can access those third party servers (hacker, insider threat) or their communications links if original content is transmitted, or where cryptographic hashes and perceptual hashes can allow linking to original content (under some circumstances).

For cryptographic hashes the content users are sharing may be identified by comparison with known hashes. For example, it would be possible to identify users sending racist memes by comparing with hashes known racist memes. The flow of a content item could be tracked by seeing where its hash turns up, even if the content is not identified. This information can also be used to identify “networks” of users sharing similar content. This could be damaging or embarrassing to some individuals.

The situation is worse for perceptual hashes which are effectively misnamed. Perceptual hashes are in fact highly compressed representations of an image, albeit normalised to a particular grid and with colour information removed. It has been demonstrated that it is possible to reconstruct images from perceptual hashes both mathematically and with the aid of AI. These reconstructed images are fuzzy and indistinct but researchers have shown that under some circumstances they may be sufficient to identify the type of content (e.g. to identify an image likely to contain nudity or sexual content) and in extremis to identify individuals.

We therefore believe that 14.24(a) and A15.18 should make at least some reference to privacy considerations otherwise we believe the code has a significant risk of inadvertently promoting use of APIs in a way that is not appropriately privacy protecting.

Companies including Cyacomb have developed and demonstrated means of matching via API which avoid these issues using privacy enhancing technologies. Approaches using homomorphic encryption have also been proposed, although we do not believe these to be practical for cryptographic hashes or available at all for perceptual hashes at this time.

The description of perceptual hashing could be interpreted to exclude the use of privacy enhancing technologies in the matching process, and we believe the wording of 14.20, 14.24(a) and/or in A15 should focus on the effect of hash matching rather than the specifics of the process, and allow for privacy enhancing technologies to sit between user content and the database.

14.39 Removing CSAM

We refer to our comments made in response to Question 1 concerning specificity of harms and responses.

This section opens by taking about "Removing CSAM" but does not make any statement about reporting CSAM. While this is required by other existing law, it would be helpful if this document at least signposted the relevant requirements and guidelines. We raise this point specifically because Cyacomb has been contacted a number of times by smaller organisations discovering CSAM on their systems and asking what they should do. It is not widely understood what is required, and the desire to "remove" promptly can result in evidence which would have been of value to law enforcement being inadvertently destroyed. There is also a lack of knowledge about how to report CSAM in smaller organisations, especially where overseas users are concerned.

14.39(a) mentions investigation, but with no clarity of who carries this out.

14.39 (d) assumes reporting to law enforcement is taking place with no information concerning how or why.

Signposting relevant information would strengthen this document.

14.42 (a) – Cryptographic Hashing

Cyacomb works with Law Enforcement on a day to day basis. In this work we see the results of both Cryptographic and Perceptual hashing and we believe that these technologies are highly complementary, but with very different characteristics.

The conclusion reached in A15.12 appears to focus on images. Video is currently much less well served by perceptual hashing approaches. Other CSAM content including sets of images that are published in a zip archive (or similar), or in document file formats including Word and PDF (with accompanying text, explanation, context) are also not currently detected by perceptual hashing.

We believe the recommendation should require platforms to assess the risk of non-image CSAM content and if that risk is significant then Cryptographic hashing should be used as well as perceptual hashing. This would almost certainly be the case for platforms that allow sharing of files, especially large video files.

The conclusion reached in A15.12 has limitations already recognised at 14.29 which recognises that prevalence is an important factor in determining the utility of a particular false positive rate. For the vast majority of services CSAM is a tiny fraction of the content available on the service.

Cryptographic hashing typically has a far lower false positive rate than perceptual hashing and can be used in these situations to create a "fast track" for content matched to the extraordinarily high degree of precision provided by cryptographic hashing where a high volume of perceptual hashing matches, many of which may be false positives, may take longer to triage. This again should be based on risk assessment considering volume, prevalence, and the needs of downstream processes.

14.42 (e) Human Moderator Review

It would be helpful to refer to law and guidelines governing this practice as Cyacomb's experience suggests this knowledge is not available to all smaller platforms.

It would also be helpful to clarify the expected sequence of events. If content has been uploaded and is identified by perceptual hashing, should the content be removed from publication presumptively prior to moderator review (then restored to publication if determined to be a false positive), or should identification through perceptual hashing be a trigger for human moderation after which confirmation the content should be removed from publication if confirmed as CSAM?

What risks and impacts should be considered in making such a decision?

14.42 (f) "or prevent from being uploaded"

Relating to comments above on 14.39, prevention of upload may also prevent effective reporting.

In Cyacomb's dealings with smaller platforms in particular, we often hear them take a view that preventing upload is desirable. This avoids the issue of having known CSAM on their platform (even in unpublished form), avoids having a moderator (or anyone else) viewing the content, and avoids any risk of accidental publication.

In working with Law Enforcement and agencies such as IWF and NCMEC we more frequently hear the view that upload should be permitted as this allows the platform to collect evidence to make a more comprehensive report to law enforcement which is more likely to be actionable.

While we recognise the law about uploading or reporting is outside Ofcom's remit, we believe that there is a risk this documentation will add to confusion unless clarification is added referring to appropriate guidance.

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.

14.73 / 14.74

See previous response related to 14.42 (f) about how a platform should determine whether removal is appropriate or reporting is required.

14.78

See previous response related to 14.24, albeit the risks of third parties and international data transfer are acknowledged here.

14.108 / 14.110

Once again refers to limitations in the current hashing provision as an influence in Ofcoms decision making. See comments in the response section above entitled "Requirement for Hashing" about the innovation ecosystem and incentives for improvement.

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.

14.140 Evidence on Hashing in CT – Gaps

From a technical perspective the false positive rates for Terrorism material can reasonably be expected to be similar to CSAM.

Cryptographic hashing will be highly accurate in matching against the database. The accuracy and completeness of content included in the database will be the determining factor of system performance. There are a number of credible commercial and NGO creators of such databases.

There has been one high profile (and to the best of our knowledge unproven) case where the integrity of a moderation database has been called into question and this related to the abuse of a CT database https://www.bbc.co.uk/news/uk-60497274. The guidance provided on database governance addresses the integrity of these databases and if followed would avoid a situation such as the one presented here (whether it actually happened or not).

In our experience working with Law Enforcement and in Online Safety perceptual hashing is less used in Terrorism applications than CSAM as content types other than images are more prevalent in Terrorism. Cryptographic hashing is perceived as a better tool to cover all content types, although there is no reason why perceptual hashing cannot be applied to images and video.

The relevance of content matching approaches (whether URL or content) is limited by the importance of context in assessing whether something is illegal in relation to terrorism. While there is some content that is illegal in itself, context is often both important and nuanced. The consequences of incorrect moderation can have serious implications for free speech in politics and religion in a way that isn't the case with CSAM.

The legality of terrorist content in different jurisdictions is also much more variable than for CSAM. A news video protected by free speech in one country could be illegal in another.

A hashing solution will match content, and ignores context at the point of creation of the hash, context at the point of matching, and jurisdictions.

For content that is in and of itself illegal automated moderation makes sense in this context exactly as it does in CSAM. This requires that a dataset (or category) of "illegal content" needs to be maintained separately to content that can or is likely to be illegal, and probably maintained independently for different jurisdictions.

For content that can or is likely to be illegal, the question of context becomes vital and human oversight is required.

We believe that automated detection remains useful and should be required (at least for large and high risk platforms). It may not always be appropriate for content to be removed from view immediately however. Hashing can contribute to identifying posts that require human review, and also to identifying high risk areas of the platform (users, groups, forums, chats) and helping a platform understand the extent to which it is being used to promote terrorist materials.

From a technology implementation perspective any platform that implements hashing for CSAM should have very low additional costs for checking a separate hash list (or hash lists) for terrorist content. Where CSAM checks are not required, the implementation costs could be expected to be similar to hashing for CSAM.

We are unable to comment on the costs of additional moderator workload or acquiring datasets as this is beyond our expertise.

Cyacomb Technology & Access to Hashes and URLs

Ofcom recognises the challenges that CSAM Hash list providers may have in scaling operations. This also applies to providers of URL lists for CSAM, or terrorism. In addition to the organisations named in the report, Cyacomb is aware of a number of commercial organisations in possession of CSAM or Terrorism hashes or URL lists that are not currently making them available externally but would in principle be willing to do so.

We see two principal barriers to hash and URL lists being available where they are needed for detection purposes.

- **Security Requirements**
    - In the UK and EU hashes are generally considered by data protection authorities to be personally identifying information as they could in principle be used to look up the hashed content and thus identify victims of abuse or their abusers. This makes them high sensitivity data under data protection regulations such as GDPR.
    - Were lists of hashes to fall into the hands of offenders they could in principle be used to access corresponding content using search engines or peer-to-peer file sharing tools. This would increase offending and lead to increased harm to survivors through their material being viewed more widely.
    - Were lists of CSAM URLs to fall into the hands of offenders they could increase accessibility of that CSAM and result in increased offending. This would also lead to increased harm to survivors through their material being viewed more widely.
    - As a result any transfer of a hash list or URL list can only happen where the supplier of that list has conducted appropriate diligence into
        - The identity of the recipient

        - The legitimacy of their need for the data
        - Their technical and governance capabilities to protect the lists
    - This diligence is costly and time consuming, as a result of which
        - The cost of diligence limits the floor price for data sharing
        - The capacity for diligence often limits the ability of an organisation to onboard new users of its lists
        - Increasing capacity for diligence requires investment in people and processes with no certainty of return.
- **Commercial Control**
    - Acquiring and maintaining lists of CSAM or terrorism content or URLs is costly.
    - Fees from providing access (or membership that gives access) to third parties for automated moderation applications is a key (and often only) source of income to cover those costs.
    - This applies whether the supplier is working on a non-profit basis and seeking to cover costs, or whether they are a for-profit commercial enterprise.
    - Transferring a hash list or URL list for the purpose of matching has a number of associated risks
        - The recipient could keep using the list after the period for which they have paid
        - The recipient could intentionally, accidentally, or as a result of a cyber security incident provide the list to third parties who may use it without paying
        - The nature of the content in lists may give a recipient an insight into proprietary processes and enable them to create their own alternative of competitive list (in order to reduce their costs or compete)
    - In each case these risks undermine the commercial value of the list and the interests of the list owner are harmed
    - This also drives the need for diligence as described above to minimise the risks.

The need for extensive knowledge of and diligence on potential users of a (hash or URL) list creates a barrier to availability, as that diligence is both costly and time consuming. Many elements of the process are hard to automate.

These requirements place a lower limit on the size and sophistication on the organisations that can gain access. To pass diligence organisations must be able to demonstrate appropriate governance and security, which may not be the case for smaller organisations. They must also be able to pay an amount that at the very least covers the cost of diligence and contributes to costs of maintaining the data.

Even if there were a large cohort of organisations that *could* pass diligence, many potential providers of lists lack the capacity to run diligence on a large number of new users in a short timescale, such as might be driven by a new regulatory requirement.

These challenges arise from the nature of current practice, which is to transfer the (hash or URL) list from the provider to the recipient organisation for matching, which in turn leads to the diligence requirement.

Cyacomb has developed technology to eliminate the need to transfer lists. It does so by packaging a list into a form called a “Contraband Filter”. Contraband Filters do not contain personally identifying information, and cannot be used to connect to or reconstruct such information. If a Contraband Filter falls into the hands of bad actors there is no useful information they can recover from it. Nothing in a Contraband Filter would let them find CSAM, CSAM

websites or terrorist materials. We describe this as being “secure by design”. Security of a Contraband Filter is designed into its nature. It is not dependent on passwords, encryption or other traditional cybersecurity measures placed around it. As such, Contraband Filters are much easier to transfer from one place or organisation to another for the purposes of matching.

Contraband Filters are compatible with both cryptographic matching and perceptual matching approaches.

Contraband filters can also be designed to incorporate licensing, such that they can only be used for the duration of time (or volume of checks) that has been paid for.

Cyacomb believes this technology could help providers of list make them widely available for automated moderation through matching in Contraband Filter form. By using such a secure-by-design approach diligence could become a much lighter touch process, scalable to a much larger number of users. We believe this would significantly reduce barriers to access to data, and costs for access as the diligence cost would be reduced. There would be a cost to using Cyacomb’s technology, in the form of a license fee or a per-usage model (Software as a Service - SaaS). The SaaS model is particularly suited to smaller organisations as it minimises integration cost (negligible) and exploits the scalability of the cloud.

SaaS models for matching have not traditionally found favour as they often results in transfer of a hash or URL to the SaaS for checking. Even in encrypted form this offers a risk to privacy. Cyacomb has developed a technology called Privacy Assured Matching that builds on the Secure by Design approach to Contraband Filters with a Privacy by Design approach to matching. Even in the matching service or communications link were compromised by a bad actor, this approach protects user privacy ensuring their content can be neither identified nor tracked through the network.

The documents under consultation do not recommend wide deployment of hash or URL matching (limiting the recommendation to the largest and highest risk services). There are indications that Ofcom would have made the requirement broader had the ecosystem been more scalable and the cost proportionate. Cyacomb believes its innovative technology can contribute to improving privacy and security, reducing the cost of diligence and adoption and enabling scale, which appears aligned with Ofcoms intent. However, with limited short term market demand and no clear indication on Ofcom on future direction there is limited incentive for commercial providers like Cyacomb to make investments in technologies such as these. We believe this is an example where Ofcom could play a stronger role through processes such as this in creating a healthy innovation ecosystem by providing more information to enable innovators to make investment decisions.

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:	