

Ofcom Additional Safety Measures Consultation – IWF response

Contents

Livestreaming – Introduction	2
Livestreaming – proposals for all users.....	3
Livestreaming – proposals to protect children	8
Proactive Technology	9
Amendments to Illegal Content Judgements Guidance for CSAM	19
Perceptual hash matching for intimate image abuse.....	20
Perceptual hash matching for terrorist content.....	22
CSAM hash-matching for pornography.....	23
Recommender Systems	28
User banning	28
User Sanctions Policy.....	32
Highly effective age assurance in the Illegal Content User-to-user Codes	33
Grooming	35
Crisis Response.....	37
Broadening Appeals	39
Combined Impact Assessment.....	39
Statutory Tests	39
Other questions	39

[Contents](#)

Livestreaming – Introduction

1. Do you have further evidence regarding the harms and risks to users from livestreamed illegal content or content harmful to children, or harms and risks to children from broadcasting livestreams?

In 2018, the IWF prepared a one-off snapshot study to directly consider the risks of livestreaming.¹ The implications of this study still apply, with livestreams remaining a high-risk functionality for the production and sharing of child sexual abuse material. Whilst IWF analysts do not engage in real-time, or near real-time, review of livestreaming content, our analysts continue to find that commercial incentives are a large driver for broadcasting CSA.

Distribution methods

As outlined in the IWF's 2018 report, memorialised videos of child sexual abuse are stored and recorded for sale on third-party services.² In the study, 100% of the images had “been harvested from its original upload location and further distributed via third party websites” (p. 14).

Offenders create a ‘grid image’ of the livestream to provide a snapshot of the type of content available for purchase. A grid image is one image depicting stills from a video in a grid format. These grid images are posted on dedicated live-streamed CSA forums. The video of the livestream is then uploaded on to a third-party file storage or file-sharing site from where it can be downloaded on payment (p. 15). This method of sharing livestreams is now viewed as an older distribution method, but remains relevant.

Rapid developments in technology have led to the emergence of other forms of distribution methods, such as websites specialising in producing ‘leaked’ images. These websites will, on request, hack into a user’s account to extract indecent images of the user. Leaked images allow offenders to access ‘new content’ and can often also be used to coerce the user for more images or financial compensation.

Financial motivators

CSEA livestreams are driven by sexual and financial motivators.³ A 2025 report from the Australian Institute of Criminology further outlines the relationship between these drivers:

¹ IWF (2018). *Trends in Online Child Sexual Exploitation: Examining the Distribution of Captures of Live-streamed Child Sexual Abuse*.

² Ibid.

³ Celiksoy, E., Schwarz, K., Sawyer, L. and Ciucci, S. (2023). *Payment methods and investigation of financial transactions in online sexual exploitation of children cases*. University of Nottingham’s Rights Lab.

[Contents](#)

“A key factor in CSA live streaming offences is the interaction between those who procure the offences and those who facilitate them (Açar 2017; ECPAT International 2017; Europol 2019; Global Alliance against Child Sexual Abuse Online 2016). Facilitators of CSA live streaming are the recipients of financial transactions procuring the abuse, who then arrange the live streaming session on behalf of the offender. However, it is important to note that, in some instances, offenders also procure CSA live streams directly from victims (Napier, Teunissen & Boxall 2021)” (p. 4).⁴

To address the financial rewards driving the livestreaming of child sexual abuse, proactive technology measures must extend to also detect for payments associated with CSEA. This intervention will be crucial in disrupting the commercial influences for producing and distributing livestreamed CSAM.

For further details on extending proactive detection duties to respond to the commercial drivers of CSEA please see our response to Question 13.

2. Do you have further evidence regarding the benefits to users or children from livestreaming?

N/A

Livestreaming – proposals for all users

3. Do you agree with our proposals? Please provide your reasoning, and if possible, provide supporting evidence

The harms presented by livestreaming have been well recorded. We welcome Ofcom’s proposal to introduce a suite of measures to tackle the risk of livestreaming CSEA.

One-to-many livestreams

However, we are concerned with these safety measures being applied only for one-to-many livestreams.

⁴ Cubitt, T., Napier, S., and Brown, R., (2025). [Financial risk indicators of child sexual abuse live streaming: A proof of concept prediction model](#). *Trends & issues in crime and criminal justice*.

[Contents](#)

Ofcom has recognised livestreaming as a high-risk functionality for CSEA and other harms to children in the Register of Risks for both Illegal harms and Children’s register of risks.^{5&6} Ofcom also notes, within the Additional Safety Measures consultation, the extensive CSEA risks of livestreaming. Considering the clear recognition of the harm this functionality poses, it is surprising that Ofcom are of the view that only a “minimum set of measures is required to set a baseline to protect users” (p. 45).⁷

Offenders can use the many-to-many livestreaming function on gaming platforms to groom children and livestreaming abuse. NCMEC recently reported a spike in offender groups contacting children on gaming platforms and getting them to livestream acts of harms against themselves and others.⁸ Core aspects of many-to-many livestreams have also been noted on the Register of Risks, including its video-sharing function and group-messaging function. Furthermore, the Illegal Harms Register of Risks also explicitly states the CSEA risk on gaming services, particularly multi-player gaming (para. 2A.27-2A.29). Despite the compounding threat that high-risk functionalities in many-to-many livestreams can pose for a child, it is concerning that these measures did not extend to all types of livestreaming.

E2EE livestreams

Gaming livestreams are increasingly becoming end-to-end encrypted (E2EE). Discord has reported that by 1 March 2026, its livestreaming feature, Go Live, will become end-to-end encrypted.⁹ Go Live is a many-to-many broadcast that allows users to share their entire screen in a Discord server, or voice call for up to 50 people at a time.¹⁰

This is uniquely important for E2EE livestreams. E2EE is another high-risk feature noted in the Illegal Harms Register of Risks. As previously mentioned, the compounding risks of harms on a many-to-many livestream significantly increase the abusability of the feature and leave children vulnerable to sizeable risk of harm.

Safety by design

Ofcom’s piecemeal approach to introducing livestreaming measures goes against the principle of safety by design. Targeting the measure to one aspect of livestreaming overlooks Ofcom’s own views outlined in the risk register.

⁵ Ofcom (2024). [Illegal Harms Register of Risks](#).

⁶ Ofcom (2024). [Children’s Register of Risks](#).

⁷ Ofcom (2025). [Additional Safety Measures Consultation](#).

⁸ NCMEC (2025). [‘Spike in online crimes against children a “wake-up call.”’](#) [web page].

⁹ Discord (2025). [‘End-to-End Encryption for Audio and Video’](#) [web page].

¹⁰ Discord (2025). [‘Go Live and Screen Share’](#) [web page].

[Contents](#)

It is crucial that Ofcom extends these measures to all types of livestreamed broadcasts to ensure every aspect of livestreaming is safer by design.

Ofcom's current proposal overlooks their own risk assessment, which recognises the harm on these services. In May 2025, the UK Government stated that they want safety by design to be "where features are chosen and designed to limit the risk of harm to users."¹¹ To meet this vision, Ofcom must be led by evidence of harm and not just evidence of interventions that work currently.

Human moderation

We are supportive of Ofcom's decision to require that human moderators are "available at any time that the livestreaming functionality is available to users" and agree with Ofcom's view that human moderators will help ensure the effective determination of content that is suspected CSA (para 5.21).

As Ofcom has noted, human moderators support the quick takedown of illegal content due to their ability to make effective judgements on suspected illegal content (para. 5.31). Human moderators are able to identify grooming behavioural patterns which require contextual understanding that extends beyond the capabilities of automated content moderation. According to INHOPE, human moderators "can detect suspected grooming behaviour by using keyword searches to scan for sexual requests within the first two days of contact."¹²

It is unclear from the consultation how the human moderator measure will apply in practice. A human moderator to be 'available' doesn't clarify how they will be able to intervene on time. This places a huge onus on the child or a user watching to report to a human moderator if there is a risk of CSA happening on the livestream. There is also a likelihood that the CSA may already be broadcast before a human moderator is able to intervene. **Ofcom needs to provide further guidance on human moderators' response and intervention duties on livestreaming. Ofcom must also clarify how human moderators will be able to intervene in E2EE livestreams.**

With many service providers cutting the number of human moderation staff, it is important that guidance sets out a proportionate amount of human moderators that all in-scope services must have.¹³ **In our view, the number of human moderators on a platform must be proportionate to the size and risk of the service. This must also account for the number of moderators needed for a dedicated CSEA reporting branch.**

¹¹ DSIT (2025). [Draft Statement of Strategic Priorities for Online Safety](#).

¹² INHOPE (2022). ['The importance of human content moderators' \[web page\]](#).

¹³ Gerken, T., (2025). ['TikTok puts hundreds of UK content moderator jobs at risk.'](#) BBC News.

[Contents](#)

For human moderation measures to be effective in tackling risks of livestreamed CSEA, Ofcom must introduce guidance on human moderation to ensure there is a minimum standard on how service providers must implement this measure. Guidance must require strong and regular training to identify risks of grooming and CSEA and robust welfare measures to safeguard the wellbeing of moderators. These requirements should equally apply for both in-house and third-party moderation teams.

Welfare

Human moderators can be exposed to highly distressing content on a daily basis and without appropriate support this can have an adverse impact on their wellbeing.¹⁴

At the Internet Watch Foundation, we have image analysts, not content moderators, but the nature of our analysts work means that, like content moderators, they are exposed daily to harmful and illegal imagery. To address the highly difficult nature of the work, the IWF's wellbeing strategy is built on care, compassion, and long-term support.

This includes:

- A trauma-informed recruitment process including meeting with counsellors to fully understand the nature of the work;
- Carefully structured training to slowly build up the skills needed to assess imagery;
- Strict schedules for image graders;
- Limiting direct exposure to child sexual abuse content to five hours a day, with hourly breaks;
- Prohibiting overtime and encouraging breaks;
- Mandatory counselling for image viewing staff; and
- An annual assessment with a clinical psychologist for image viewing staff.

The IWF's approach is continues to evolve, shaped by clinical advice, external audits, and team feedback.¹⁵

4. Do you agree with our assessment of the impacts (including costs) associated with this proposal? please provide any relevant evidence which supports your position.

We are encouraged to see the assessment of the direct financial costs for service providers in introducing livestreaming and human moderation measures takes a balanced approach. As noted by Ofcom, setting up a reporting category will be a one-off cost. Whilst introducing

¹⁴ Kleinman, Z., (2024). 'I was moderating hundreds of horrific and traumatising videos.' BBC News.

¹⁵ IWF (2025). 'How we protect our people.' IWF Annual Data & Insights Report 2024.

[Contents](#)

human moderation will be an ongoing cost, as Ofcom have noted, these costs are proportionate to the size of the service providers in scope.

It is important for Ofcom to balance the short-term business costs of implementing these measures alongside with the long-term social, safety, and economic benefits these measures will create for UK users. Recent research by the NSPCC and Baringa shows that a business model centred around child safety online (CSO) can be profitable.¹⁶ The found that introducing a CSO business model has the potential to be an important influence on consumer choice equal to that of consuming organic produce and sourcing sustainable goods.

5. Do you have any views on the optimal design of reporting functions and choice categories for users to report content that depicts the risk of imminent physical harm? Include any evidence, such as, testing to optimise wording, design of tools to support users to submit accurate and timely reports and how these may be used to support moderation actions.

It is encouraging to see a specific reporting category for imminent risk of physical harm, including threats to life and CSEA. Having a dedicated line for reporting CSEA should ensure an immediate response to impending risks of CSA being broadcasted.

Ofcom states that the value of a livestream is reduced if the broadcast depicting a violent attack or CSEA is taken down faster (para. 5.36). This is because “incomplete footage is less appealing and is less likely to be circulated as widely (as set out in para. 5.32 and 5.33) -meaning a decreased impact in terms of, for example, radicalisation” (para. 5.37, p. 50). We would disagree with this assessment for CSAM, as even a short period of broadcasted CSA results in the generation of multiple images of child being sexually abused.

An immediate response to cases of CSEA is essential to reduce the risk of CSAM being broadcasted. **To achieve this, users must have clear access to a dedicated reporting button for specifically child sexual abuse and exploitation.** As Ofcom have noted, additional reporting flow will not confuse users and make them less likely to report. Instead, a targeted complaint button will ensure take down of a livestream that could produce multiple images of CSAM in a few seconds alone (p. 48).

To reduce the risk of services reacting slowly to a broadcast of CSA, all services must have a dedicated 24/7 emergency response team for CSEA. **Moderators must also have the power to end a livestream where there is confirmed broadcast of CSAM.**

¹⁶ NSPCC and Baringa (2025). [‘Child Safety Online: Good for business opportunity’ \[PowerPoint presentation\]](#).

[Contents](#)

This risk could also be remedied by amending the crisis response protocol to include CSEA as a priority illegal harm. Please see our response to Question 49 for further details.

6. Do you consider that there are alternative measures which would materially reduce the risks to users from livestreaming such as preventive safety by design frictions, prompts or restrictions? If so, please detail them and provide evidence on the costs and efficacy.

For our response on safety by design frictions, please see Question 3.

Livestreaming – proposals to protect children

7. Do you agree with our proposals? Please provide your reasoning, and if possible, provide supporting evidence.

Screen capture

It is positive that Ofcom is proposing users are unable to screen capture a child's livestreams, including through third party tools where technically feasible. The IWF sees the dangerous outcomes of livestreaming daily. In 2024, IWF confirmed 291,273 reports of CSAM. 91% of the reports assessed as criminal were found to contain 'self-generated' imagery.¹⁷

However, we do not see the validity in limiting the screen grab measure, using third party tools, to only when it is technically feasible. As Ofcom has recognised, "there are several screen capture and recording prevention tools commonly used by video streaming, banking and some messaging services" (para. 6.29). Whilst there may be complexity in implementing these tools, as Ofcom have noted, this should not be conflated with technical infeasibility. Placing a technical feasibility condition clips the ambitions of an important child safety measure.

Ofcom must provide guidance on how they will be measuring when it is technically feasible for a service to implement their measures, and when it is not.

Ofcom must also provide case studies to inform platforms about how they would assess technical feasibility for separate features and functionalities, including regularly updating the guidance in line with new and emerging technologies.

¹⁷ IWF (2025). [IWF Annual Data & Insights Report 2024](#).

[Contents](#)

For further details on our concerns with the technical feasibility condition please see our response to Question 11.

Gifting and commenting

We welcome the proposal to require gifting and commenting to be turned off for livestreams by a child and recognise this will be a valuable child safety feature. These features are incentives which can encourage children to perform sexual acts. One example in the IWF's livestreaming snapshot study found:

“One child, who gave her age as 12 years old, referred to having 50 viewers to her current broadcast stream. After repeatedly exposing herself to the webcam, she stated that she would stop the broadcast if people didn't start commenting or “liking” the stream as there would be “no point” in her continuing” (p. 13).¹⁸

8. If you are a service provider, what measures do you currently undertake to moderate livestreams and protect children who undertake livestream broadcasts, and what is your evidence on the effectiveness of such measures?

N/A

9. Do you consider that there are alternative measures which would materially reduce the risks children face when livestreaming, both in general and in relation to operation of the supporting functionalities of comments, reactions, gifting and content capture? If so, please detail them and provide evidence on the costs and efficacy.

N/A

10. Do you agree with our assessment of the impacts (including costs) associated with this proposal? please provide any relevant evidence which supports your position

Please see our response to Question 4.

Proactive Technology

¹⁸ IWF (2018). [Trends in Online Child Sexual Exploitation: Examining the Distribution of Captures of Live-streamed Child Sexual Abuse](#).

[Contents](#)

11. Do you agree with our proposals? Please provide your reasoning, and if possible, provide supporting evidence

We strongly welcome proactive detection measures extending to a wider range of CSAM, including disguised or altered CSAM URLs and CSAM discussion (para. 9.22 and para. 9.23). Widening the scope of CSAM detection will be important for identifying where child sexual abuse is taking place. In particular, proactive detection of CSAM discussion will be an invaluable tool for enabling the identification and takedown of CSAM. For the IWF, keyword detection is a foundational part of how Analysts work quickly to identify CSAM and action a report. We also provide a Keyword List for our members to help them identify and take down illegal content on their own services.¹⁹

Proactive detection of disguised CSAM URLs will be another important lever in tackling online CSA. The IWF has found that CSAM remains live for longer on sites with disguised and altered URLs due to their hidden nature. In 2024, the IWF uncovered 2,877 disguised websites sharing CSAM. Access to these sites is granted via a link on a different website, requiring a specific cookie or optional HTTP referrer. For IWF analysts, accessing and taking action on disguised sites typically takes at least twice as long as it does for an average website.²⁰

Top-level Domain (TLD) Hopping

When a CSAM URL is identified, platforms should also detect for the same URL using other common top-level domain names (e.g., “.net”) to reduce the risk of domain hopping.

Disguised websites continue to exploit top-level domain hopping to avoid detection and remain online, making them recognisable to offenders. The IWF maintains a Top-level Domain (TLD) Hopping List that records second-level domains of sites dedicated to child sexual abuse. As of 2024, this list contains 104 unique second-level domains and has tracked 296 individual historical hops.²¹

End-to-end encryption

Where communication is moved to end-to-end encrypted (E2EE) messaging, this increases the abusability of the feature. As Ofcom has previously noted, E2EE presents clear risk factors for child sexual abuse.²² The NSPCC’s 2025 research outlined that limited restrictions in direct messages between and adult and child users allow immediate opportunity to begin grooming and sharing harmful content.²³

¹⁹ IWF (n.d.) ‘[Keywords](#)’ [web page].

²⁰ IWF (2025). ‘[Commercial disguised websites](#).’ IWF Annual Data & Insights Report 2024.

²¹ IWF (2025). ‘[Top-level domain hopping](#).’ IWF Annual Data & Insights Report 2024.

²² Ofcom (2023). [Protecting people from illegal harms online consultation: Volume 2](#).

²³ NSPCC and PA Consulting (2025). [Targeting girls online: how online services enable the abuse and harassment of girls](#).

[Contents](#)

However, as highlighted in the IWF’s recent explainer *Preventing the upload of child sexual abuse material (CSAM) in E2EE environments* (2025),²⁴ the upload prevention method can tackle these risks via pre-checks for known CSAM *before* the content is sent through E2EE messaging. In fact, as the explainer outlines, this method is already used on E2EE platforms to detect for malware and generate rich preview links (see pages 9 - 10).

The IWF defines upload prevention as:

“[A] safety feature that stops harmful or illegal content such as CSAM from being shared before it is sent. Upload prevention compares a file’s hash against a trusted hash list, and if a match is found, the upload is blocked at source. This process can take place entirely on-device or through privacy-preserving lookups” (p. 25).²⁵

In short, when a hash-match of known CSAM is found, the content cannot be uploaded and sent through E2EE messaging.

Ofcom must require all service providers to adopt the upload prevention method on all public communication channels, including those that are end-to-end encrypted before an image or video is sent.

Technical feasibility of upload prevention

Ofcom risks undermining these positive measures by raising the technical feasibility condition. The current interpretation of “technical feasibility” offers services a loophole to avoid their safety duties, and undercuts safety measures which are crucial in the fight against online child sexual abuse.

We welcome Ofcom’s decision to investigate where a service provider claims that it is technically infeasible for them to implement proactive technology. Yet, it remains unclear how Ofcom will assess a service’s claim of technical infeasibility and when an investigation would be launched. As our 2025 explainer highlights, the upload prevention method is technically feasible to implement, given the method is already deployed on E2EE services to detect for content that is not CSAM.

The OSA Network’s technical feasibility brief (2025) explains the barriers Ofcom’s interpretation of the technical feasibility condition continue to present for providing effective safeguarding measures, including those outlined in this consultation.²⁶ We are supportive of the OSA Network’s call for Ofcom to provide stronger clarification between technical infeasibility and cost. Further, Ofcom should conduct regular review of the technology to ensure the technical infeasibility argument is not redundant. When it is technically feasible for a service to

²⁴ IWF (2025). [Preventing the upload of child sexual abuse material \(CSAM\) in E2EE environments](#).

²⁵ Ibid.

²⁶ Woods, L., and Walsh, M., (2025). [Requirements for codes measures - technical feasibility and proportionality](#). OSA Network.

[Contents](#)

implement a safety measure, Ofcom must amend the codes to reflect the change. Finally, it is essential that services cannot introduce a “roll-back” on their functionality so it no longer technically feasible for them to implement a measure.

12. Do you have any comments on the Proactive Technology Draft Guidance?

These views are supported by the Centre for Emerging Technology and Security (CETaS)

We welcome Ofcom’s proposal to have service providers implement proactive technology to detect for CSAM, should it be technically feasible. Requiring proactive technology to be assessed and reviewed, in line with the proactive technology criteria, will ensure there is a minimum standard for tools deployed for proactive detection of CSAM. Whilst Ofcom clearly note that hash-matching is a form of proactive technology (see p. 132 of the consultation), the proposed proactive technology criteria have been designed with the review of artificial intelligence (AI) tools in mind.

We recommend the following amendments to ensure criteria are technology-agnostic, adaptable, and robust:

- Require data sets to be both trusted and of high quality
- Implement a layered set of detection tools
- Adjust the evaluating performance criteria to consider both AI and hash-matching technology
- Introduce criteria for privacy intrusiveness and security

Trusted and high-quality data

The expectations of high-quality data criteria should apply equally to providers sourcing data sets from third parties and those developing and/or updating in-house data.

We agree with Ofcom’s definition of high-quality data as stated on page 9, Annex 6 of the consultation:

““High quality” refers to data that is accurate; legally and ethically sourced; well-labelled (where appropriate); representative of the harm and context the proactive technology is intended to address; and sufficiently diverse to support meaningful evaluation, including (where relevant) to test the proactive technology’s ability to detect content it has not previously encountered.”²⁷

High quality data should be accurate, diverse, appropriately sourced and representative of the targeted harms. However, it is also important that the data are trusted, so there is both transparency and integrity regarding how the data are processed. For example, in a trusted

²⁷ Ofcom (2025). [Additional Safety Measures Consultation](#).

[Contents](#)

dataset, such as those prepared by the IWF, the provider should be able to justify why and how each piece of content has been added.

The IWF plays a vital role in providing the internet industry with several quality-assured datasets to prevent the spread and upload of known CSA images and videos. These include image hashing utilising Microsoft’s PhotoDNA, a URL blocking list of live webpages, and keywords list. Key to this is our trusted relationship with the internet industry which enables us to act as a broker between them and government and law enforcement.

Good data collection practices will be needed where there are clearly defined thresholds for what is or is not CSAM, coupled with a clear process on how that judgment is made at each stage of the process. It is essential for any trusted data set to have a secure method for protecting highly sensitive data like CSAM. **We encourage Ofcom to amend the data criteria to factor not just the quality of the data, but also whether it is trusted. Where these data sets are developed in-house, it is essential there is a third-party review of the data set by an authoritative body.**

Implementing a layered set of detection tools

Ofcom has proposed that providers use proactive technology to detect harms that closely align with the eight targets for immediate action. One of these targets includes preventing the sharing of child sexual abuse material, as outlined in Ofcom’s 2024 Online Safety Act progress update:

“Offenders can’t share child sexual abuse content and children don’t face unsafe contact. We want the sharing of CSAM to become much more difficult, and children to be protected from contact from adults they don’t already know. Almost a fifth of children experience sexual solicitation from adults they have chatted with online – change is needed to stop this” (p. 8).²⁸

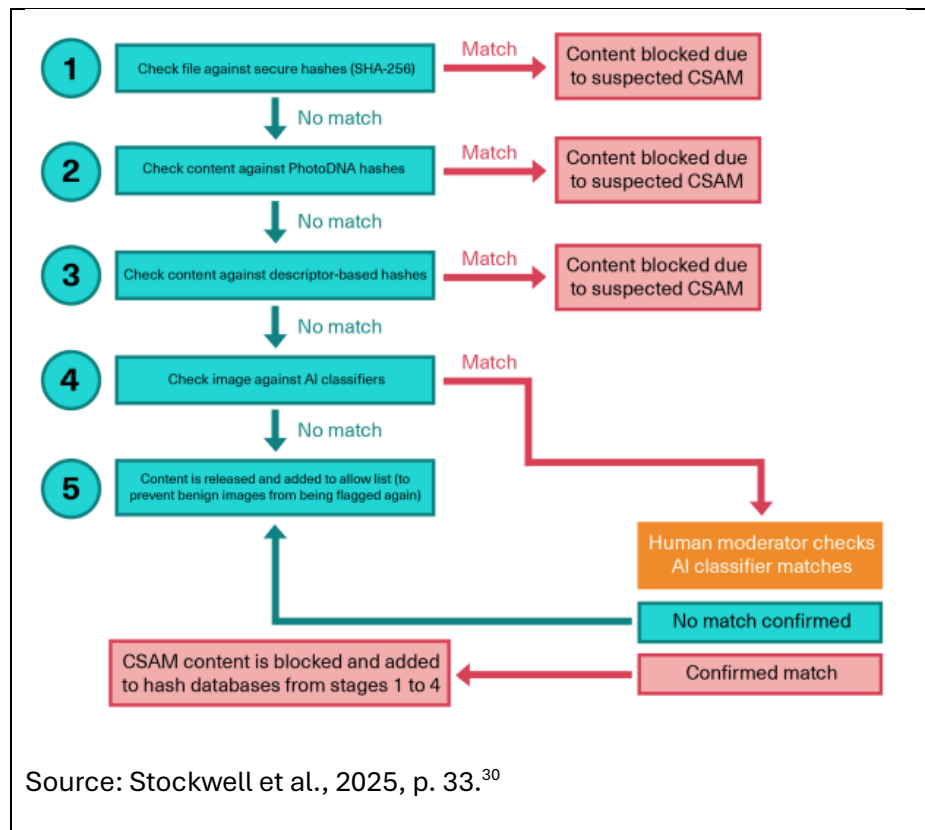
To make the sharing of CSAM “much more difficult”, services must deploy a layered set of tools which can work together to detect both known and unknown CSAM. This can include hash matching technologies, AI based tools, and human moderation.²⁹ Figure 1 provides an overview of the layered moderation process.

Figure 1: Overview of a layered moderation process for CSAM matches

²⁸ Ofcom (2024). [Implementing the Online Safety Act: progress update.](#)

²⁹ Stockwell, S., Wake, G., Dargahi, T., Ajao, O., Latham, A., Abdullahi, A. D., and Sexton, D. (2025). [Privacy-preserving Moderation of Illegal Online Content.](#) Centre for Emerging Technology and Security.

Contents



Cryptographic and perceptual hashing are highly effective ways of detecting known image-based child sexual abuse material. The IWF’s hash list uses both types of hashes (Microsoft’s PhotoDNA, MD5, SHA-1 and SHA-256) to detect for confirmed images and videos of child sexual abuse.³¹ Layered upon hash-matching, AI-based tools can detect suspected CSAM on a platform. Tools such as Thorn’s Safer Predict have classified nearly 2 million files of potential CSAM.³² However, human evaluation will be essential in confirming whether the suspected content does depict child sexual abuse. Once a human moderator has reviewed the suspected content and confirmed whether it is or is not CSAM, the content can be added to the services hash list to detect for future cases of now known CSAM. The content classification by the human reviewer will also support reinforcement learning to improve the AI tool’s recall and false positive rates.

Evaluating performance for AI based detection

³⁰ Ibid.

³¹ IWF (n.d.). ‘[Image Hash List](#)’ [web page].

³² Thorn (2024). ‘[Introducing Safer Predict: Using the Power of AI to Detect Child Sexual Abuse and Exploitation Online](#)’ [blog].

[Contents](#)

‘Precision and recall’ are an appropriate and necessary metric when assessing the efficacy of AI based detection tools. Research, particularly in the higher education sector, continues to highlight the persistent issue of AI detection tools producing false positives.³³ When extended to CSAM detection on online platforms, this challenge is further amplified because of the volume of content reviewed using AI. As Chad M.S. Steel (2024) notes:

*“The base rate problem in CSEM detection appears to present an insurmountable challenge for AI-based approaches at a provider level. Because the prevalence of CSEM is so low on most legitimate social media platforms when compared to non-offending content, even a 99% detection accuracy results in unmanageable false positive rates” (p. 5).*³⁴

The appropriate balance between precision and recall for AI classifiers will depend on a case-by-case scenario. However, we would not agree with Ofcom’s view that human review should apply to an “appropriate proportion of the outputs”³⁵ (p. 82). Rather, human review should be proportionate to the likelihood of false positives.

Evaluating performance for hash matching

Hash matching is recognised as an industry standard method of detecting CSAM due to its high accuracy rates and low false positives. Where there are false positives, these fall into one of two categories: technical or human. Trusted organisations like the IWF mitigate the risk of false positives entering the data set by requiring human review of the data (see ‘trusted and high-quality data’ for further information).³⁶

The current phrasing of the evaluation criteria does not appropriately capture how to evaluate hash-matching technologies. Performance evaluation for hash matching must consider both human errors and technical errors. By technical errors, we mean a ‘hash collision’ when a confirmed image-based CSAM matches against a different image. However, the likelihood of a collision is almost negligible when using modern cryptographic hashes. Concerns about performance evaluation can be further minimised by services demonstrating transparency and integrity in how they source their data.

Privacy intrusiveness and security

In the consultation, Ofcom has defined proactive technology as “rules-based software such as hash matching or keyword detection, Automated content classifiers (algorithmic systems which

³³ Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., and Waddington, L. (2023). [Testing of detection tools for AI-generated text](#). *International Journal for Educational Integrity*.

³⁴ Steel, C. M. S. (2024). [Artificial intelligence and CSEM - A research agenda](#). Child Protection and Practice.

³⁵ Ofcom (2025). [Additional Safety Measures Consultation](#).

³⁶ IWF (2025). [Hotline assessment overview](#). IWF Annual Data & Insights Report 2024.

Contents

analyse content and associated signals), and other proactive technologies which analyse inputs – including network analysis, user profiling and behavioural analysis” (p. 72).

A report by the Centre for Emerging Technology and Security (CETaS) shows that there are also a wider range of promising techniques that could assist in detecting illegal content but may be at risk of falling outside of this defined scope.³⁷

These include cryptographic-based methods such as Zero-Knowledge Proofs (ZKPs) and Private Set Intersection (PSIs), which are designed to enhance the privacy-preserving nature of hash-matching processes. Both methods are particularly valuable on end-to-end encrypted environments. They offer an additional of security for datasets layering on top of hash-matching techniques. This enables service providers to detect whether user content matches a database of known harmful material without sharing full content or disclosing any personal data (see pages 39-52).³⁸

To promote a more comprehensive evaluation criteria, Ofcom should draw on CETaS’ research which has highlighted how additional metrics that focus on privacy intrusiveness and security robustness are available. These include:

- **Data exposure metrics – these criteria would cover the extent to which user data are accessed, processed and retained by different actors.**
- **Cryptographic protection metrics – these criteria would ensure the robustness of cryptographic safeguards as part of the moderation process on E2EE networks.**

13. Do you agree with the harms currently in scope of these measures? Are there any additional harms that these measures should capture? Please provide the underlying arguments and evidence that support your views, including evidence regarding the availability of accurate and effective proactive technology.

Confidential sections have been redacted.

We welcome the fact that the proposed measures cover illegal harms including image-based CSAM, CSAM URLs, grooming, and content harmful to children.

However, these measures should also extend to electronic payment service providers to tackle the financial drivers of generating and distributing CSAM.

Electronic payment service providers

³⁷ Stockwell, S., Wake, G., Dargahi, T., Ajao, O., Latham, A., Abdullahi, A. D., and Sexton, D. (2025). [Privacy-preserving Moderation of Illegal Online Content](#). Centre for Emerging Technology and Security.

³⁸ Ibid.

[Contents](#)

Commercial incentives have a strong influence on the generation and distribution of CSEA.^{39, 40 &}

⁴¹ For many service providers, hosting CSEA is a core part of their income generation.⁴²

Preventing payments from being completed disrupts the cycle of financial incentives fuelling the generating and sharing of child sexual abuse material ('commercialisation of CSAM'). The IWF continues to find a steady growth in the commercialisation of child sexual abuse material. Between January and June of this year, there was a 72% increase in confirmed cases of sexual extortion, with many cases financially motivated.⁴³

Invitation Child Abuse Pyramid (ICAP) sites sell CSAM using a pyramid-style link-sharing model. IWF Analysts have observed that ICAP sites continue to operate on the clear web in large numbers without any attempts to conceal the CSAM they are selling. This means that these sites are visible to the public. [Redacted].

The Treasury's 2025 National Risk Assessment of Money Laundering and Terrorist Financing directly names online CSEA as a form of money laundering threat.⁴⁴ The risk assessment states:

"The proceeds from online child sexual exploitation and abuse are criminal funds as soon as funds are sent in return for access to child sexual abuse material or livestreams, or in the case of child sexual extortion as soon as the victim has been manipulated into sending funds. Victims of extortion can also be used for money laundering purposes as money mules. **Electronic Money Institutions and Payment Service Providers are regularly used to make payments associated with online child sexual exploitation and abuse, although retail banking, cryptoassets, and other payment methods are also used.** There is no global estimate for the proceeds generated by online child sexual exploitation and abuse. Often, individual transactions or payments are small in value; however, this is in stark contrast to the very high level of harm caused to victims"⁴⁵ (p. 26, emphasis added).

³⁹ Cubitt, T., Napier, S., and Brown, R., (2025). [Financial risk indicators of child sexual abuse live streaming: A proof of concept prediction model](#). *Trends & issues in crime and criminal justice*.

⁴⁰ Celiksoy, E., Schwarz, K., Sawyer, L. and Ciucci, S. (2023). [Payment methods and investigation of financial transactions in online sexual exploitation of children cases](#). University of Nottingham's Rights Lab.

⁴¹ Childlight (2025). [STUDY B: Where does the money flow? An exploratory study on the financial structures of organised crime linked to child sexual exploitation](#). Searchlight 2025.

⁴² Salter, M. and Richardson, L. (2021). [The Trichan takedown: lessons in the governance and regulation of child sexual abuse material](#). *Policy & Internet*.

⁴³ IWF (2025). ['Child sexual extortion cases in the UK soar with warnings 'ruthless' criminals are still putting children and young people at risk'](#) [news].

⁴⁴ HM Treasury and Home Office (2025). [National Risk Assessment of Money Laundering and Terrorist Financing 2025](#).

⁴⁵ Ibid.

[Contents](#)

Ofcom is currently proposing that financial illegal harms that are in scope of proactive detection duties extend to fraud (para. 9.21). It is also important financial services are not being abused for criminal activity. **To tackle the commercial aspects of CSAM, these duties must also extend to the digital financial transactions around CSEA.**

Where proactive technology has identified image, URL, and/or text-based child sexual abuse material, this must trigger the technology to then detect for any accompanying payment processor linked to the webpage. Where a payment processor has been detected, an alert must be sent to the financial service so they can act on the suspected report.

The IWF offers this service through our Payment Band Alerts to protect legitimate financial services from being used for payments for online criminal child sexual abuse imagery. The IWF Hotline takes quick action to notify services if their company name or brand is used to pay for child sexual abuse images or videos. This includes sharing the associated email addresses and merchant IDs to ensure swift take down of the content.

Whilst there is pre-existing guidance to tackle terrorist financing and anti-money laundering, the same does not extend for child sexual abuse material.⁴⁶ **With HMT listing online CSEA as a human trafficking offence (see above), Ofcom must work with the Financial Conduct Authority (FCA) to produce guidance on how financial services must respond to their services being used for the selling and purchase of child sexual abuse material.**

Proactive detection for IIA

We would encourage Ofcom to consider extending proactive technologies duties to IIA as well. The ambitions of Ofcom's IIA hash matching measure ICU C14 are to "work in tandem with the existing moderation measures" in the Illegal Harms Codes (para. 11.15). Ofcom also notes in the consultation that ICU C14 "complements existing automated content and search moderation measures that tackle child sexual abuse material (CSAM) in the Illegal Content User-to-User and Search Codes" (para. 11.17).

14. Do you agree with who we propose should implement these measures? Are there any other services that should be captured for some or all of the relevant harms?

We welcome the fact that these measures will apply to all file-storage and file-sharing services, regardless of size, all user-to-user (U2U) services at high risk of grooming, and all large services

⁴⁶ FCA (2024). [Payment Services and Electronic Money – Our Approach \(version 6\)](#).

[Contents](#)

at medium risk of grooming. It is a needed first step to apply these measures to high risk U2U services with more than 700,000 monthly active users.

However, given the high-risk nature and severity of the harms these measures are proposing to tackle, we believe the threshold should be reduced to ensure more service providers come into scope. This is important as we know that CSEA harm can occur on high-risk smaller services. The 700,000 thresholds should be lowered to ensure more small but risky sites are in scope of the proactive detection duties, particularly those with a high portion of child users.

We are also concerned that the technical feasibility condition presents E2EE services with the opportunity to circumvent rolling out proactive technology to detect for CSAM on their platforms. As we stated in Question 11, it is technically feasible to check for CSAM before content is encrypted and Ofcom must take particular to investigate claims around technical infeasibility.

15. Do you agree with our assessment of the impacts (including costs) associated with this proposal? please provide any relevant evidence which supports your position.

Please see our response to Question 4.

Amendments to Illegal Content Judgements Guidance for CSAM

16. Do you agree with our proposal? Please provide your reasoning, and if possible, provide supporting evidence.

It is positive to see that metadata will be used to determine whether CSAM is being shared in publicly communicated environments, including E2EE environments. Although the IWF Hotline's proactive detection does not extend to E2EE environments, our Analysts often encounter links to E2EE group chats on the clear web that plainly depict and/or suggest CSAM being shared or sold on the group chat. The new illegal content judgment guidance will be important in this content being taken down.

However, Ofcom risks undercutting these positive measures through the technical feasibility condition. We know it is technically feasible for service providers to detect for content before it is encrypted. These principles are used to detect for malware on an encrypted platform or generate rich preview links. Instagram's nudity detection tool also follows the same principles. The current interpretation of "technical feasibility" offers services a loophole to avoid their safety duties, and undercuts safety measures which are crucial in the fight against online child

[Contents](#)

sexual abuse. We discuss our concerns on the technical feasibility condition further in Question 11.

17. Do you have any evidence relevant to the examples given?

Confidential sections have been redacted.

[Redacted].

Proactive detection, using services like the IWF's Keyword Detection List will be essential to help detect E2EE messaging where CSAM is being shared or distributed. **Additionally, Ofcom must demand that E2EE services go further and implement the upload prevention method on their devices to check for known image based CSAM and URLs before the content can be shared.**

18. Do you agree with our assessment of the impacts (including costs) associated with this proposal? please provide any relevant evidence which supports your position.

Please see our response to Question 4.

Perceptual hash matching for intimate image abuse

19. Do you agree with our proposals? Please provide your reasoning, and if possible, provide supporting evidence.

We welcome Ofcom's decision to extend perceptual hash matching to detect intimate images and videos which are non-consensually shared online. As Ofcom has noted in the consultation, the sharing of non-consensual intimate images (NCII) is both widespread and increasing.

NCII requires a determination of non-consent, which is not feasible using only the image - unlike CSAM, where all material involving children is non-consensual. A possible solution to the issue of identifying material as NCII may involve having a similar expert body to the IWF, but with the additional step of a determination of non-consent by a civil court. Once an absence of consent is determined, the material could be hashed and entered into a central registry of NCII, which platforms can then use to block or remove it - mirroring the IWF's successful model for CSAM.

[Contents](#)

To effectively tackle the spread of NCII, such content must be placed on equal footing with CSAM and treated as illegal content. This will mandate that service providers are required to block the image.

20. Do you have any evidence on the relative efficacy of third-party and internal databases for image-based IIA content?

The IWF is extremely proud of the model we have established, which helps technology companies prevent the upload, storage, and sharing of child sexual abuse imagery online, while preserving the privacy of its users. Whilst the IWF does not find, assess, or remove NCII content, we share technology and lessons learned from our CSAM work to strengthen the ecosystem around NCII. Crucially, the technology itself is agnostic, meaning it can be applied to different types of harmful content, including NCII.

The IWF and SWGfL have worked in the past to identify, remove, and hash material created by offenders. A key example of how this content can be tackled together is the role the IWF and SWGfL played in finding, removing and hashing (creating digital fingerprints) child sexual abuse imagery created by jailed online sexual predator, Abdul Elahi. Our models are precise and privacy-protecting since they enable companies to scan for illegal content without processing or storing the actual images. Hash-matching is also an approach which has proven to be scalable, effective, and trusted by industry.

21. Do you consider this measure to be effective for file-sharing and file-storage services? Please explain your reasoning and, if possible, provide supporting evidence.

We believe applying this measure to file-sharing and file-storage sites will be highly important to tackling IIA. Ofcom has already noted in the consultation that these services are being abused for IIA (footnote 344, p. 146), thus there is a clear need to address the harm of IIA on these services.

Applying perceptual hash-matching duties on file-sharing and file-storage services has been crucial in tackling CSAM online. Notably, this has allowed Ofcom to launch investigations against services suspected to be non-compliant which has prompted said services to take appropriate action.⁴⁷ We look forward to the same enforcement powers extending to IIA.

⁴⁷ Ofcom (2025). [‘Enforcement programme into measures being taken by file-sharing and file-storage services to prevent users from encountering or sharing child sexual abuse material \(CSAM\)’](#) [web page].

[Contents](#)

22. Do you agree with our assessment of the impacts (including costs) associated with this proposal? please provide any relevant evidence which supports your position.

We will defer to the assessment of the South West Grid for Learning on the impacts associated with this proposal. We believe that any impact assessment by Ofcom should balance the short-term business costs of implementing these measures alongside with the long-term benefits to society. Further information on our views can be found on Question 4.

23. Do you consider this measure to be effective for large general search services? Please explain your reasoning and, if possible, provide supporting evidence

We will defer to the expertise of the South West Grid for Learning on the effectiveness of this measure for large general search services.

Perceptual hash matching for terrorist content

24. Do you agree with our proposals? Please provide your reasoning, and if possible, provide supporting evidence.

N/A

25. Do you have evidence regarding the accuracy and effectiveness of hash matching solutions for detection of terrorism content specifically (including their false positive and false negative rates).

N/A

26. Do you have evidence on the extent to which a hash matching solution can identify terrorism content accurately when applied in different contexts from that in which the hash was created, noting the potential implications for freedom of expression.

N/A

27. Do you have a view on the degree of human oversight required to support the use of hash matching in relation to terrorism content?

N/A

[Contents](#)

28. Do you have evidence or views on the impact assessment (including costs) associated with implementing and maintaining hash matching technology for the detection of terrorism content (such as the impacts and costs of setting up an internal database, connecting to an external provider, and moderation costs).

N/A

CSAM hash-matching for pornography

(AKA Proposal to extend perceptual hash matching for child sexual abuse content)

29. Do you agree with our proposals? Please provide your reasoning, and if possible, provide supporting evidence.

Confidential sections have been redacted.

The IWF recognises that adult pornography is legal in the UK. There is an important opportunity and pressing need to work with the adult industry, and other companies which host a significant amount of pornographic content if we are to build better protections to prevent the spread and normalisation of CSAM. It is imperative that online pornography sites prevent child sexual abuse on their platforms, and IWF can provide tools to help make this possible.

We welcome Ofcom's proposal to extend the scope of the CSAM perceptual hash-matching measure to all pornography providers, regardless of size. As Ofcom has noted, pornography services are understood to be a high-risk area for the distribution and upload of CSAM and are one of the most common types of platforms where CSAM is found on the clear web (para. 13.4-13.5). IWF Analysts frequently report finding CSAM on pornographic services. They also find pornographic services being used as a disguise for criminal websites showing CSAM ('disguised sites'). Baroness Bertin's Independent Pornography Review (2025) highlights further information on disguised sites:

"Evidence provided by the IWF highlighted the use of adult pornography as a disguise by criminal websites that are used to display child sexual abuse material. These sites show child sexual abuse material when accessed through banners on other sites in the same commercial groups. However, when accessed directly, these websites will only show adult pornography (sometimes with young looking adult performers) to prevent the site being identified as criminal. The disguises are often very convincing. The disguised pages may allow you to view adult images and videos on the site, and links on these pages will lead to

[Contents](#)

other adult pornography sites – and you will not be able to access child sexual abuse images from the disguised version of a site. The huge number of adult pornography sites on the internet makes this an effective disguise to fool visitors who may be unaware of the existence of hundreds of commercial child sexual abuse sites and will not be aware that these sites are hiding the most appalling imagery” (p. 109).⁴⁸

Pivotal to preventing the upload of child sexual abuse content online is the IWF Hash List, which is a growing list of nearly three million hashes of child sexual abuse images and videos. The Hash List includes hashes of imagery that features children aged 14 to 17. This is important due to the increased risk that the children in these images may be mistaken for adults, and imagery of their abuse could then be uploaded to adult content websites.

We believe this proposal will appropriately target the risk of image-based CSAM on pornographic services. However, we feel that Ofcom can go further with this measure to support deterrence from re-offending. **After a hash-match has been found, this should trigger pornographic providers to share awareness raising resources, like the reThink chatbot, to deter individuals from seeking CSAM whilst providing supportive resources to access help.**

As Ofcom has noted, our reThink chatbot partnership with the Lucy Faithful Foundation and Aylo led to significant statistical decrease in the number of searches for CSAM on Pornhub UK.⁴⁹ The evaluation found that warning messages did not only reduce the number of CSAM searches, but also lowered the likelihood of re-searching for CSAM.⁵⁰ Paired with hash-matching duties, we believe that deterrence messaging and signposting will be an important remedy in discouraging repeat offending. We further discuss the importance of awareness raising messaging as a deterrence measure in our response to Question 36.

Image Intercept

The IWF has also recently introduced a powerful new safety tool, Image Intercept, for free to smaller platforms to prevent the spread of known CSAM using our hash-matching technology.⁵¹ This tool has been made possible through funding from the UK Home Office.⁵² Image Intercept allows small businesses, online platforms, and start-ups to use our database with a simple code plug-in. We believe this tool will protect a raft of sites from child sexual abuse images and videos. [Redacted].

⁴⁸ Baroness Bertin (2025). [Creating a safer world: the challenge of regulating online pornography](#). Department of Science Innovation and Technology.

⁴⁹ Scanlan, J., Prichard, J., Hall, L. C., Watters, P. A., and Wortley, R. (2024). [reThink Chatbot Evaluation](#).

⁵⁰ Walsh, M., Denis, D., and Findlater, D. (2023). [Deterring online child sexual abuse and exploitation: lessons from seven years of campaigning](#). *The Faithfull Papers*.

⁵¹ IWF (2025). [‘Image Intercept: Simple, Effective Protection Against Online Child Sexual Abuse Imagery’](#) [web page].

⁵² IWF (2025). [‘Teens face ‘crisis’ of online sexual exploitation as charity says major new Government-backed scheme will ‘put rocket boosters’ on Online Safety Act’](#) [news].

[Contents](#)

IWF's Standard of Good Practice for Adult Services

In 2023, the IWF declared a partnership with Aylo (formerly MindGeek) to develop a Standard of Good Practice for Adult Content Providers.⁵³ The Standard outlines essential criteria that adult services should meet to have confidence that imagery of child sexual abuse imagery is prevented. To forge this path, the IWF assembled an Expert Advisory Board to spearhead the development of these standards, comprising experts from academia, law enforcement, and civil society.⁵⁴

Following this work, the IWF's tools and services that prevent the upload, storage, and sharing of child sexual abuse material are now available to companies that specialise in publishing adult content through our membership model.⁵⁵

Electronic payment service providers

It is necessary for Ofcom to consider the influence of digital payments in raising the standards of pornographic platforms. For instance, in 2020, Mastercard made the decision to cut ties with Pornhub.com after the network's review of the website uncovered illegal content.⁵⁶ Visa and Discover similarly axed their acceptance of their cards on the site pending the completion of their own investigations. In turn, Pornhub raised the standard of its services, including ending uploads from unverified users and banning the ability of users to download much of the site's content.

The power of payment companies to enforce change in the standards of adult services is clear. It is important that Ofcom works with the FCA to produce guidance on how financial services must respond to their services being used for the selling and purchase of child sexual abuse material.

We discuss the importance of tackling the financial drivers behind the distribution and generation of CSAM further in Question 13.

Whilst the scope of this measure focuses on tackling CSAM on pornographic services, we think it is important to also note the emergence of CSAM on AI chatbots.

AI chatbots

We are very concerned that the development of AI chatbots and the ease, speed, and quality at which text-to-image generative AI (Gen AI) tools have been developed is enabling a range of harms relating to child protection, including an avenue for offenders to generate AI CSAM, gain

⁵³ IWF (2023). [‘Model of good practice for adult sector to be developed by IWF and MindGeek’](#) [news].

⁵⁴ IWF (n.d.). [‘Adult Content Standards in Combating Online Child Sexual Abuse Imagery’](#) [web page].

⁵⁵ Ray-Hill, D. (2025). [‘IWF working with the adult sector is vital if we’re serious about tackling child sexual abuse imagery online’](#) [blog]. IWF.

⁵⁶ Bell, K. (2020). [‘Pornhub ends unverified uploads and bans downloads’](#) [news].

Contents

information about how to commit CSEA, or to simulate the offence of sexual communications with a child.

For the first time, analysts at the IWF have identified AI-generated child sexual abuse images that are connected to AI chatbots.

The IWF received 17 reports of CSAM on chatbots between 1 June 2025 and 7 August 2025. The reports contained mostly (94%) Category C imagery and featured predominantly 11–13-year-olds (82%). One report depicted a 7 to 10-year-old child. The reports concerned one website, which the IWF was alerted to following two public reports. The website, marketed as an “AI girlfriend” website, featured fantasy AI companions which contained non-photographic actionable imagery, as well as photo-realistic AI CSAM.⁵⁷ The photorealistic child sexual abuse imagery was found if a user clicked on a chosen chatbot from a preview page on the website showing the different chatbots. Examples of the scenarios that users could ‘play out’ with the chatbots included ‘child prostitute in a hotel’; ‘sex with your child while your wife is on holiday’; and ‘child and teacher alone after class’.

Users have the option of generating more images from the one that has already been created and the library of chatbots was created by the website admin, as well as website users – meaning this site would likely qualify as a user-to-user service and be in scope of the Online Safety Act. The categories provided different chatbots for different interests, including 2D, 3D, male and female.

While AI-generated child sexual abuse imagery is illegal under UK law, there is a lack of clarity regarding how chatbots fall under scope of the Online Safety Act

The OSA is clear that “any AI-generated text, audio, images or videos that are shared by users on a user-to-user service is user-generated content and would be regulated in exactly the same way as human-generated content” (para. 6).⁵⁸ Therefore, if illegal content such as CSAM has been created by AI, service providers in scope of the Act still need to act on it. The Act also requires that when service providers change their regulated services significantly - for example by adding a new AI chatbot as part of the regulated service - they are required to do a fresh risk assessment and take any necessary mitigating measures.

Ofcom has previously set out how the Online Safety Act will apply to Generative AI and chatbots.⁵⁹ Ofcom’s letter emphasises that it will not treat content created by AI tools and chatbots differently than content created by human users for purposes of the Act. Any text, images, or videos created by these chatbots will be considered “user-generated content” and regulated by the OSA. However, in April 2025, Ofcom noted that, with regards to the challenges

⁵⁷ Hardy, E. (2025). ‘[AI chatbots and child sexual abuse: a wake-up call for urgent safeguards](#)’ [blog]. IWF.

⁵⁸ Ofcom (2024). ‘[Open letter to UK online service providers regarding Generative AI and chatbots](#)’ [news].

⁵⁹ Ibid.

[Contents](#)

presented by Gen AI and the rise of chatbots, the legal position was complex and not entirely clear.⁶⁰

Professor Lorna Woods has identified different types of chatbots and how they may be covered by the regulation. She has highlighted that “the coverage appears incomplete and there are some technical questions which remain unanswered which may affect the completeness of protection.”⁶¹

We would welcome action by Ofcom to clarify positioning on AI chatbots and companions under the Online Safety Act and the legal requirements that in-scope platforms are required to comply with. We also request that Ofcom provide an update on its work to consult industry on how risks relating to Gen AI chatbots can be mitigated.⁶²

Furthermore, we believe there are a number of steps that providers should be taking to mitigate harm. This includes:

- To prevent the upload, generation and sharing of explicit images of children.
- To introduce safeguarding measures for chatbot and companion products, such as pop ups and automated responses to seek support when users share words indicative of abuse.⁶³
- To employ classifiers to identify specific types of content in the model’s responses and filter sensitive content.
- Enforce minimum age requirements on all chatbot applications and websites to protect children online.
- For search services to de-index websites which host AI companions that generate AI CSAM.

30. Do you agree with our assessment of the impacts (including costs) associated with this proposal? please provide any relevant evidence which supports your position.

⁶⁰ Parliament. House of Commons (2025). [Oral evidence: Social media, misinformation and harmful algorithms](#). (HC 441).

⁶¹ Woods, L. (2025). ‘[Chatbots and the Online Safety Act](#)’ [web page]. OSA Network.

⁶² In April 2025, Ofcom’s Director for Online Safety Strategy Delivery, Mark Bunting, was asked how Ofcom was responding to the challenges presented by Gen AI during a Select Committee oral evidence session in Parliament. On chatbots, he said that the “legal position” was “not entirely clear” and “complex.” He noted the regulator would be consulting with industry to understand how risks could be mitigated. It would be helpful for Ofcom’s findings to be made public.

⁶³ UNICEF (2022). [Safer Chatbots Implementation Guide](#).

[Contents](#)

Ofcom notes that the main costs associated with the implementation of this proposal are one-off. These include building and maintaining the hash list, cost associated with the purchasing of hardware, image assessment, and reporting duties (para. 13.8). However, as outlined above, the IWF's Image Intercept tool can support smaller pornographic content providers to comply with the new measures at no additional cost.

Recommender Systems

31. Do you agree with our proposals? Please provide your reasoning, and if possible, provide supporting evidence

N/A

32. Do you have evidence on what types of content are typically recommended to users as part of concerted foreign interference activity?

N/A

33. Do you have evidence on whether services track the extent of algorithmic amplification, such as impressions and reach, of content that is later deemed illegal/violating. If so, do they (or does your service) use this information to enhance the safety of their systems?

N/A

34. Do you agree with our assessment of the impacts (including costs) associated with this proposal? please provide any relevant evidence which supports your position

N/A

35. Are there any impacts of the proposed measure that we have not identified? Please provide the rationale and any supporting evidence for your response

N/A

User banning

[Contents](#)

36. Do you agree with our proposals? Please provide your reasoning, and if possible, provide supporting evidence.

We welcome Ofcom's proposal to introduce bans for adult users who share, generate or upload CSEA. We consider it a proportionate response to introduce a ban for all types of CSEA, rather than just imposing a sanction. This is an effective approach to tackling repeat offending and protecting children from being exposed to known offenders on a platform. **We also welcome further clarity from Ofcom on how they will ensure service providers take reasonable steps to prevent users from returning to the platform.**

As Ofcom noted in the consultation (para. 16.16), all types of CSEA have an adverse impact on victims and survivors. Additionally, offending is often connected to multiple CSEA harms, starting from grooming a child to the production of child sexual abuse material and/or engaging in sexual activity offline. Extending the ban to all types of CSEA harms also disrupts offending patterns and prevents the risk of offenders finding loopholes to shift tactics.

We believe Ofcom is taking the right approach by extending the ban beyond the platform the user has been found sharing, generating, or uploading CSEA. It is appropriate the user is banned from all other platforms operated by the service and unable to access any third-party service that is registered with the account's single service sign on.

However, there may be some instances where a user has unwillingly been the recipient of CSAM. It is important on those occasions that there is an appeals function available for the user to take appropriate recourse against the ban.

Cross-platform signal sharing

However, there remains a risk of offenders moving to platforms that fall outside the above two categories. **To ensure safety by design, the service provider that has introduced the ban on the user account has a duty to share the relevant information with other services to prevent the offender from creating accounts on other services to engage in CSEA activity.** It is technically feasible for such information to be shared by service providers. The Lantern Programme is an example of signal sharing systems already being used on a voluntary basis by service providers to tackle the spread of CSEA.⁶⁴

Deterrence messaging and signposting

Banning offenders from services will be an important tool in keeping children safe online. **To strengthen the ambitions of this measure, deterrence measures must also consider the importances of awareness-raising messages to educate and inform a user about the consequences of their action and to signpost them to support.** The Lucy Faithfull

⁶⁴ Tech Coalition (n.d.). '[Lantern: Advancing child safety through signal sharing](#)' [web page].

[Contents](#)

Foundation's (LFF) StopItNow service provides an anonymous [helpline, and email and chat](#) services for anyone with concerns about child sexual abuse or people seeking information on how to prevent it. This has led to positive behaviour changes amongst un-arrested people who have offended.⁶⁵ For those making follow-up calls with LFF's Advisers or using their self-help modules, there were sustained positive behaviour changes. Without providing appropriate explanations for why a service is blocking an account, and appropriate signposts to resources that will support positive behaviour changes, there is a risk that offenders move to darker parts of the internet.

[Upload prevention](#)

Finally, services must also implement upload prevention on their platforms to ensure that the user is unsuccessful in sharing and uploading CSEA content, before they are blocked from the service.

For further details on the upload prevention method see Question 14.

37. What is your assessment of the options we set out in relation to the treatment of child users and which option do you consider to be most appropriate? Please provide any supporting evidence to support your arguments.

There is a need for service providers to implement interventions targeting children sharing CSAM. Placing a permanent ban on child user, in the first instance, is not the appropriate response in tackling this problem and we are concerned by the punitive approach being presented.

Imposing an immediate and permanent ban on a child user forms an underlying assumption that in every instant a child has shared CSEA content *willingly* and with *intent* to cause harm. This view overlooks the various reasons why a child is sharing such content which can range from consensually sharing images as part of an age-appropriate relationship, to being the result of coercion, extortion, and trafficking. As outlined in the 2023 report by the CSA Centre:

“There are a wide range of reasons why children share imagery of themselves. Most commonly, it is voluntarily shared – for example, as a way of indicating that they like someone, for fun, or as part of a sexual relationship (Katz and El Asam, 2020; Baines, 2018). **However, some children, particularly girls, feel pressure to send images, or do so because they worry the other person will lose interest in them if they don't (Revealing Reality, 2022a). Around one in five children share images because of coercion, sexual extortion, or trafficking (Baines, 2018).** Research suggests that,

⁶⁵ Walsh, M., Denis, D., and Findlater, D. (2023). [Deterring online child sexual abuse and exploitation: lessons from seven years of campaigning](#). *The Faithfull Papers*.

[Contents](#)

when images of children are obtained and used for extortion, boys may be more likely than girls to be extorted for money (Canadian Centre for Child Protection, 2022) while girls are commonly extorted for more images (Baines, 2018). **This reinforces the importance of recognising each individual’s motivation for and experience of both sharing and receiving images”** (p. 4, emphasis added).

The report also explains other aspects which place certain children at risk of sexual harm online. These factors include being a victim of sexual abuse, precarious family and living conditions (e.g., separation from the family, parental incarceration and parental substance abuse), gender, special educational needs and disabilities, low mental health, sexuality, and social isolation.⁶⁶ It is unclear why banning child users was determined to be proportionate response to deter children from sharing CSEA, despite children’s unique vulnerabilities being acknowledged within the consultation (para. 16.7).

The impacts of this punitive approach can be jarring for children, particularly those who are victims and survivors, and can adversely impact a child’s sexual development. According to Yang et al. (2024), “individuals who have experienced sex guilt seem to have less satisfactory sex lives and diminished sexual well-being” (p. 999).⁶⁷ There is also risk that children who are banned from regulated sites may move to darker, unregulated sites where they may be exposed to more extreme CSEA content. It is crucial these interventions are developed in consultation with child development experts to ensure they do not cause unintended long-term negative impacts for children.

Platforms will need to take a case-by-case approach for responding to children sharing, generating, or uploading CSEA. These approaches should begin by providing: (1) age-appropriate awareness building resources, (2) supportive messaging similar to Lucy Faithfull Foundation’s rethink Chatbot or IWF’s ‘Think Before You Share’ campaign, and reporting tools such as the IWF and NSPCC’s Report Remove. It is vital that the messaging is developed in a child-friendly format that avoids victim-blaming language. Research by the CSA Centre (2023) indicates that education programmes have worked in preventing harmful sexual behaviour amongst children. Their report also outlines what aspects should be included in messaging targeted for children around sexual behaviour online (p.7).⁶⁸ A 2023 report by the Policing Institute for the Eastern Region (PIER) and the IWF found that a core prevention method in

⁶⁶ Brown, S. and Tregidga, J. (2023). [Key messages from research on harmful sexual behaviour in online contexts](#). Centre of Expertise on Child Sexual Abuse.

⁶⁷ Yang, T. A., Dolan, A., Hernandez, V., Kaufman, A., Kruk, M., Robbins, K., & Conley, T. D. (2024). [Associations between negative sexual messaging in childhood and sex guilt in adulthood](#). *Analyses of Social Issues and Public Policy*.

⁶⁸ Brown, S. and Tregidga, J. (2023). [Key messages from research on harmful sexual behaviour in online contexts](#). Centre of Expertise on Child Sexual Abuse.

[Contents](#)

protecting girls from requests for self-generated indecent imagery was to raise awareness through meaningful discussions with the child and carer, coupled with effective monitoring.⁶⁹

Such upstream interventions should be followed by temporary bans and/or strikes against child user accounts found to be engaging in ongoing CSEA activity. We believe permanent bans should only be used as a final resort in cases where the account continues to actively persist with such activity, despite repeat warnings and temporary bans.

38. Do you agree with our assessment of the impacts (including costs) associated with this proposal? Please provide any relevant evidence which supports your position

N/A

User Sanctions Policy

39. Do you agree with our proposals? Please provide your reasoning, and if possible, provide supporting evidence

We are broadly supportive of measures that exist to deter the sharing and selling of CSAM and believe these deterrence measures are a proportionate response for adults engaging in illicit activities. However, we would discourage similar sanctions extending to children as a first response. Please see our response to Question 36 for further detail.

We are supportive of Ofcom introducing measures to act against generating, uploading or sharing illegal content, and content harmful to children. In the case of generating, uploading or sharing CSEA, we note that Ofcom has proposed to enforce bans as it is the most severe form of sanction. We comment on Ofcom's proposed user banning measures for adults and children in our responses to Questions 36 and 37 respectively.

We believe sanctions are a necessary deterrence measure and a proportionate response to adults engaging in illicit activities. However, like our view on banning adult users from a service, we encourage user sanction policies to be supplemented with awareness-raising campaigns and signposting. Awareness-raising messaging is an invaluable tool in encouraging long-term positive changes in offending habits. Our response to Question 36 provides further detail on the value of incorporating awareness raising messaging as an intervention for in-scope services.

⁶⁹ Daly, E., Lundrigan, S., Conway, P., Fjellidal, M., and Devendran, P. (2023). [Talk Trust Empower](#). Policing Institute for the Eastern Region.

[Contents](#)

40. Do you agree with our assessment of the impacts (including costs) associated with this proposal? please provide any relevant evidence which supports your position.

Please see our response to Question 4.

Highly effective age assurance in the Illegal Content User-to-user Codes

41. Do you agree with our proposals? Please provide your reasoning, and if possible, provide supporting evidence.

Establishing the age of a user, and whether the user is a child or young person (under 18), is foundational to the success of measures which ensure age-appropriate experiences and prevent exposure to illegal or inappropriate content.

We particularly support the requirement for platforms to ensure children cannot access pornography, and welcome Ofcom’s decision to extend CSAM hash-matching to all pornography providers (see our response to Question 29). Exposure to this content at an inappropriately early age can normalise harmful sexual behaviours, and is increasingly linked to growing sexual violence against girls and young women. Research conducted by the Children’s Commissioner’s Office in 2024 found that some groups of children feel particularly exposed to pornographic content, as stated by a 16-year-old girl⁷⁰:

*“Things like porn should be banned or **massively limited in some way**. It’s bad enough that adults watch it, but now children are being exposed to the harmful culture. As most porn videos include at least one act of violence towards women in some way which negatively impacts the way young children, especially boys, will see girls and women” (p. 22, emphasis added).*

Additionally, age assurance enables platforms to tailor user experiences and take proportionate measures relating to the design or operation of the service to effectively mitigate and manage the risks of harm to children in different age groups. IWF had previously raised concerns about a “one-size-fits-all” approach to under 18s, noting the importance of age-appropriate experiences for different age groups. Following a consultation on the draft measures, we were pleased to see

⁷⁰ Children’s Commissioner (2024). [“I’ve seen horrible things”: children’s experiences of the online world](#).

[Contents](#)

the regulator take on board IWF's call for age-appropriate access to content, features, and functionalities, by adjusting the content moderation and recommender settings.

We have noted that primary driver for this suite of measures is to define and codify highly effective age assurance (HEAA) in the Illegal Harms and Protection of Children's Codes. We welcome this decision as it will address provide clarity and consistency on implementation, particularly on how services must consider usability, fairness and privacy when determining the age of a user.

To protect the privacy of the user, it is important that age verification technology and methods do not use more personal data than as part of the age verification process. We believe that **expanding the criteria for HEAA to also incorporate the principles of data minimisation and reusability from the ICO's Age-Appropriate Design Code will help achieve this goal.**

By adopting the principles of data minimisation and reusability, the risk of services holding additional information about a child is reduced. Concurrently, it ensures the content a child is exposed to, over their time on the platform, is age appropriate.

Whilst one of the four principles of implementing HEAA is to ensure providers consider the ICO's Age-Appropriate Design Code (AADC), it remains unclear how services are expected to meet the requirements of the HEAA Guidance and AADC. There must be greater clarity on the regulatory coordination between the HEAA Guidance and AADC to ensure effective implementation of the age assurance measures. This is particularly where age assurance methods will be used to provide supportive resources on high-risk functionalities like livestreaming (i.e., ICU F3).

We will defer to the views of 5Rights for a more detailed analysis on accreditation of age verification tools.

42. Do you agree with our proposal to introduce age assessment appeals measures into the Illegal Content User-to-user Codes (ICU D15 and D16)? Please explain your reasoning.

N/A

43. Do you agree with our proposed amendments to codify the definition of highly effective age assurance in the Protection of Children User-to-user Code? Please explain your reasoning.

N/A

[Contents](#)

44. Do you agree with our proposed amendments to the Part 3 Highly Effective Age Assurance Guidance? Please explain your reasoning.

N/A

45. Do you agree with our assessment of the impacts (including costs) associated with this proposal? please provide any relevant evidence which supports your position.

Please see our response to Question 4 for our view on balancing the short-term upfront industry costs against the long-term positive outcomes for UK users. We also defer to the views of 5Rights and the NSPCC for a further analysis on the associated impacts.

Grooming

AKA Increasing effectiveness for U2U settings, functionalities and user support.

46. Do you agree with our proposals? Please provide your reasoning, and if possible, provide supporting evidence.

Confidential sections have been redacted.

We agree with Ofcom’s proposal to remove the existing provision that in-scope services should apply ICU F1 and F2 to the extent that they have existing age assurance means. In line with Part 3 Highly Effective Age Assurance Guidance (HEAA), it is important that ICU F1 and F2 should apply to all in-scope services.

The consequences of grooming are severe. In 2024, the IWF confirmed 291,273 reports of child sexual abuse imagery with 29% of the images and videos hashed depicted Category A child sexual abuse.⁷¹ [Redacted].

Establishing the age of a user, and whether the user is a child or young person (under 18), is foundational to the success of measures which ensure age-appropriate experiences and prevent exposure to illegal or inappropriate content. Further consideration should also be given to ensuring adults are not able to pose as children, which we know is often a tactic used by offenders to groom children into sharing ‘self-generated’ imagery.

The NSPCC’s 2025 report found that once a child’s email and phone contacts were imported onto a child’s account, they were being encouraged to connect with adults who owned these

⁷¹ IWF (2025). [IWF Annual Data & Insights Report 2024](#).

[Contents](#)

accounts (p. 26).⁷² Some platforms in the study also deployed questionable methods to encourage a child to connect with the adult account (i.e., obscuring the ‘decline’ option).

As the report outlines, having the phone or email details of a person does not equate to a trusted contact:

“In their interviews, technology experts highlighted a significant flaw in the design principles employed across online platforms – namely, the assumption that a user’s connections are inherently trustworthy. This fails to acknowledge that children can make connections expediently, without much thought, and on the basis of a false sense of trust (for example, engendered through mutual connections). **It also does not mirror social relationships in the physical world where trust is built over time and constantly assessed**” (p. 23, emphasis added).

To reduce the pernicious risks of grooming, ICU F1 must be amended so only the child can request to connect with an adult account. This must also include preventing an adult from connecting with a child’s account even where there is pre-existing information on the child user (i.e., email account). This would make it harder for adults who are not trusted by the child to be able to contact them.

Additionally, where an adult user is connected with a child account, this should not immediately guarantee the adult’s ability to send a direct message (DM). The NSPCC suggests two interventions to help achieve this. First, services must require that adults are not able to send a child a DM unless the child has specifically confirmed the adult is a trusted user.⁷³ Second, DMs should not be available between user accounts that have newly connected. Instead, messages sent during this time should be blurred or hidden. In circumstances where the content may have suspected nudity, the child user account must be informed of the suspected content.

47. Do you agree with option A and option B in increasing the effectiveness of the ICU F1 and F2 measures?

As mentioned above, it is good that the highly effective age assurance (HEAA) duties will extend to all services in-scope of ICU F1 and ICU F2. Applying HEAA to determine whether a user is a child or an adult will be crucial in ensuring that the safety defaults are tailored for child users.

Internet Matters 2025 data briefing found a significant gap in children reporting the negative experiences they have online. Their research found that 35% of children reported being contacted by a stranger, 29% reported pornographic content, and only 17% reported nude or

⁷² PA Consulting and NSPCC (2025). [Targeting Girls Online: How online services enable the abuse and harassment of girls](#).

⁷³ Ibid.

[Contents](#)

semi-nude images of another child (p. 5).⁷⁴ A little over half the children surveyed (54%) believed that platforms had reporting processes that were clearly and simply written (p. 10).⁷⁵ It is vital that the support information available for child users is designed to be accessible for children. **We believe that Option A is the suitable way to implement these measures. By extending these safety prompts - intended for child users - to all users, there is a risk that the information is not tailored to children and fails to be appropriately absorbed by child users.**

Additionally, as Ofcom have correctly noted with the benefits of Option A:

“[I]f service providers used highly effective age assurance to determine which of their users are adults, materially more children would benefit from the protections afforded by ICU F1 and F2. This would reduce the amount of grooming that takes place online, thereby delivering significant benefits.”

Where repeated attempts are being made by an adult account to contact unknown child accounts, service providers must investigate the account and take appropriate actions against it.

48. Do you agree with our assessment of the impacts (including costs) associated with this proposal? please provide any relevant evidence which supports your position.

Please see our response to Question 4.

Crisis Response

49. Do you agree with our proposals? Please provide your reasoning, and if possible, provide supporting evidence.

We disagree with (1) the narrow definitions of a crisis that Ofcom has proposed to incorporate into the Illegal Harms Codes and Protection of Children's Codes, and (2) the harms in scope of this measure.

The definitions prioritise mitigating the spread of threat over the severity of the threat. In its current form, the definitions overlook an important opportunity to build in safeguards against the imminent threat of child sexual abuse being facilitated both online and offline. As Ofcom

⁷⁴ Internet Matters (2025). [Understanding & improving how children report online harm: Data briefing](#).

⁷⁵ Ibid.

[Contents](#)

has acknowledged, “illegal and/or content harmful to children stemming from crises poses an **imminent risk to people both online and offline**” (para. 20.6, emphasis added).

The *Online Safety Act* (2023) defines illegal content that is a ‘priority offence’ as: (1) terrorism offences (Schedule 5), (2) CSEA offences (Schedule 6), and (3) other priority offences such as threats to kill, public order offences, harassment, stalking and fear or provocation of violence, and foreign interference (Schedule 7). The proposed harms in scope of ICU C15 target Schedule 5 and Schedule 7 offences so it is unclear why this does not extend to Schedule 6 offences (CSEA).

Introducing child sexual abuse and exploitation as a priority illegal harm in scope of this measure will be a vital recourse in protecting children from imminent risk of abuse and exploitation. Ofcom’s proposal in the Consultation to establish a direct line of communication between the service provider and law enforcement to facilitate fast and reliable information sharing would be a particularly valuable intervention to target offending and prevent contact offences (para. 20.48).

50. Do you agree with our proposed definition of ‘crisis’? Please explain your reasoning, and if possible, provide supporting evidence.

Please refer to our response on Question 49.

51. Do you consider these measures to be effective for services that are not large services? Please provide any evidence on the role of services that are not large services during crises.

N/A

52. Is there any evidence of best practice in responding to a crisis that we have not identified? Please explain your reasoning, and if possible, provide supporting evidence.

N/A

53. Do you agree with our assessment of the impacts (including costs) associated with this proposal? please provide any relevant evidence which supports your position.

N/A

[Contents](#)

Broadening Appeals

54. Do you agree with our proposals? Please provide your reasoning, and if possible, provide supporting evidence.

N/A

55. Do you agree with our assessment of the impacts (including costs) associated with this proposal? please provide any relevant evidence which supports your position

N/A

Combined Impact Assessment

56. Do you think our package of proposed measures is proportionate for services in scope of the Illegal Content User-to-User Codes, taking into account the existing package of measures, the impact on reducing the risk of relevant harms and the implications on different kinds of services?

N/A

57. Do you think our package of proposed measures is proportionate for services in scope of the Protection of Children User-to-User Code, taking into account the existing package of measures, the impact on reducing the risk of relevant harms and the implications on different kinds of services?

N/A

Statutory Tests

No specific questions

Other questions

Not about proposed measures or in the [main document](#).

For more information, please contact the Policy and Public Affairs team: ppa@iwf.org.uk

[Contents](#)

58. In relation to our equality impact assessment, do you agree that some of our proposals would have a positive impact on certain groups? Please explain your reasoning and provide supporting evidence where possible.

N/A

59. Do you consider that our proposals could have any negative impacts on certain groups? If so, please explain your reasoning.

N/A

60. In relation to our Welsh language assessment, do you agree that our proposals are likely to have positive, or more positive impacts on opportunities to use Welsh and treating Welsh no less favourably than English? 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.

N/A