Illegal Harms Consultation – NSPCC Response

Volume 2: The causes and impacts of online harm

Ofcom's Register of Risks

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

We agree with Ofcom's assessment of the causes and impacts of online child sexual abuse and exploitation (CSEA). Below we highlight further evidence which should inform Ofcom and services' assessments of these harms. We would also like to direct Ofcom to the evidence review recently published by NSPCC, focusing on children's exposure to online sexual risks and the role technology plays in exacerbating or reducing these risks.[1] Drawing from studies in the UK and internationally, it provides an up-to-date assessment of online child sexual abuse.

Prevalence

Online grooming has reached unprecedented levels. The most recent data shows that almost 34,000 online grooming crimes have been recorded by UK police since 2017/18.[2] In 2022/23, 6,350 Sexual Communication with a Child offences were recorded – an increase of 82% since the offence first came into force in 2017/18. These offences are increasingly impacting younger children. Over 5,500 offences over the last six years were targeted against primary school children, with under-12s impacted in a quarter of cases. Similarly, child abuse image offences are at the highest level on record. In 2021/22, 30,925 offences were recorded by UK police, an increase of 66% in five years.[3] Childline has also seen an increasing number of calls about online sexual abuse and exploitation, with the first half of this year seeing a 46% increase in these calls compared to the same period last year.[4]

Online CSEA is facilitated across a wide range of online services, with 150 different apps, games and websites recorded by UK police as being used to groom children online.[5] However, the risk has been concentrated on the biggest platforms. According to the most recent police data[6], Snapchat was involved in 26% of online grooming offences and 43% of child abuse image offences. After Snapchat, the offences are concentrated on Meta-owned products (Facebook, Instagram, and WhatsApp), with 47% of online grooming offences and 33% of child abuse image offences taking place on these platforms.

Impact

We would like to reinforce that the impact of online sexual abuse on children is no less than the impact of offline abuse. NSPCC research shows that these forms of abuse share many of the same consequences for children, including self-blame, depression, trouble sleeping, and self-harm.[7] The use of technology can also lead to additional psychological effects. For example, if the abuse has

[1] Bryce, J. et al (2023) Evidence review on online risks to children. London: NSPCC.
[2] NSPCC (2023) 82% rise in online grooming crimes against children in the last 5 years.
[3] NSPCC (2023) We're calling for effective action in the Online Safety Bill as child abuse image crimes reach record levels.
[4] Childline data from April-September 2023 (Q1-Q2 2023/24).
[5] NSPCC (2023) 82% rise in online grooming crimes against children in the last 5 years.
[6] NSPCC (2023) We're calling for effective action in the Online Safety Bill as child abuse image crimes reach record levels; NSPCC (2023) 82% rise in online grooming crimes against children in the last 5 years.
[7] Hamilton-Giachritsis, C. et al (2017) "Everyone deserves to be happy and safe": A mixed methods study exploring how online and offline child sexual abuse impact young people and how professionals respond to it. London: NSPCC.

involved image-sharing, victims and survivors are often fearful that the images will be shared with others in the future, or that individuals will try and find them based on the images.

Victims of online CSEA must receive appropriate professional support.[8] As part of their duty to mitigate the impact of harm experienced on their service, regulated platforms should play a role in providing funding for victim-survivor support and recovery services. In the future, we recommend Ofcom redistributes part of any income generated from enforcement fines for breaches to CSEA duties to these vital services.

"Sometimes they'll be, you know, **literally nightmares of him trying to stab me or kill me or strangle me** or something like that, generally, they're nightmares, but **the flashbacks, on a day-to-day basis, tend to happen if there's a trigger**. It can be something so small sometimes."
Quote from a young person who was groomed online – NSPCC research.[9]

"**There's evidence of it and I don't know who else has seen that… It does make me feel a bit sick.** I don't know, it does stress me out. I think it makes me anxious because **I don't know what was recorded**, when it was recorded."
Quote from a young person who experienced online CSEA – NSPCC research.[10]

Experience of girls

Online CSEA disproportionately impacts girls. Whilst Volume 2 notes the gendered nature of this harm, it is consistently caveated by the fact that boys are likely to under-report and so this may skew the data on this issue. Whilst this may be the case, and the recent evidence on financial sextortion has shown how significant the impact for boys can be, the evidence is clear. Girls are the target in four in five grooming cases[11], and last year the Internet Watch Foundation (IWF) reported that girls were shown in 96% of the online child sexual abuse material (CSAM) they saw, compared to 2% for boys[12]. Research with young people has found that girls are considerably more likely to receive explicit images from others without consent, which is supported by calls to Childline – in one year, around 70% of the reports to Childline about receiving a sexually explicit image were made by girls. Our recent evidence review also finds that the impacts of online sexual harassment and intimate image abuse tend to be greater for girls.[13]

The data summarised here is from a range of sources, including children's reporting, law enforcement, proactive detection, and research. It consistently indicates that girls are particularly impacted by technology-assisted CSEA, and this should be reflected in Ofcom's final publications.

The disproportionate harm which girls experience online was recognised in the Act by requiring Ofcom to produce best practice guidance on protection for women and girls. As this guidance will in part be based on the Codes of Practice, it is vital that the experiences of girls online are appropriately recognised and that there are strong measures to tackle violence against women and girls (VAWG). The specific experiences of girl online, which are shaped by both their age and gender,

[8] Bryce, J. et al (2023) Evidence review on online risks to children. London: NSPCC.
[9] Hamilton-Giachritsis, C. et al (2017) "Everyone deserves to be happy and safe": A mixed methods study exploring how online and offline child sexual abuse impact young people and how professionals respond to it. London: NSPCC.
[10] Hamilton-Giachritsis, C. et al (2017) "Everyone deserves to be happy and safe": A mixed methods study exploring how online and offline child sexual abuse impact young people and how professionals respond to it. London: NSPCC.
[11] NSPCC (2023) 82% rise in online grooming crimes against children in the last 5 years.
[12] IWF (2023) IWF Annual Report 2022: #BehindTheScreens.
[13] Bryce, J. et al (2023) Evidence review on online risks to children. London: NSPCC.

as well as many other factors, are often overlooked in an effort to only look at age *or* gender; we urge that Ofcom recognises the intersectional nature of online harms.

There is a gap in the evidence-base regarding how grooming and CSAM impact non-binary children or those with other gender identities.[14] Ofcom's evidence review should recognise that this group of children is currently overlooked in research and data, and more work needs to be done to understand the experiences of this group of children.

Question 2: 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.

Grooming

We agree with the risk factors that have been identified. However, this section would be significantly strengthened if the most impactful risk factors were highlighted. By highlighting the risk factors which have the greatest impact on children's safety, services will be better placed to carry out more nuanced risk assessments and target the riskiest design features. For example, a service may only have one relevant risk factor, but if this has been identified as a key risk for CSEA then they should still be considered medium-high risk and ensure mitigating the risks posed by this design feature is a priority.

Evidence overwhelmingly shows that end-to-end encryption / private messaging, image sharing capabilities, cross-platform communication, and the ability for perpetrators to create fake profiles are key risk factors for online grooming and this should be recognised in Volume 2.

The evidence behind these risks is outlined in more detail in answer to Q16 and our recent evidence review.[15] Whilst there are other important risk factors, these factors are consistently exploited by perpetrators to groom children. Coercing children to send nude images and then using these for blackmail and exploitation, typically over private messaging, is particularly prominent.

CSAM

We broadly agree with the analysis in this section but note that AI-generated CSAM is not referenced in the review of how CSAM offences manifest online. There is significant evidence to show that AI is being used to generate CSAM. The IWF have reported that their analysts are removing Category A (the most serious type) AI-material, and that they have discovered online manuals which help offenders to write prompts and train AI to produce increasingly realistic images.[16]

The proliferation of AI-generated CSAM poses a major risk to children's safety. This imagery normalises the sexual abuse of children. It is also often uses real images of children, victimising and re-victimising these children every time these images are generated and viewed. Moreover, as the technology advances and the availability of these images grows, it will become increasingly challenging for police forces to identify the children who are being impacted and need urgent protection.

[14] Bryce, J. et al (2023) Evidence review on online risks to children. London: NSPCC.
[15] Bryce, J. et al (2023) Evidence review on online risks to children. London: NSPCC.
[16] IWF (2023) How AI is being abused to create child sexual abuse imagery.

The scale at which AI-generated CSAM can be reproduced and altered poses a major and complex challenge to detection and moderation. Platforms must understand the nature and scale of this harm, as well as its impact on children.

CSEA – Virtual Reality

There should be greater recognition of how grooming and CSAM offences manifest in virtual reality environments. Virtual reality (VR) poses a number of risks to children's safety. Children have experienced 'contact' abuse, grooming and exploitation in VR environments.[17] Perpetrators have used immersive technologies to enact child sexual abuse on 'virtual' children; these children are sometimes 3D model depictions of real-life children, such as child actors or children known to the offender.[18] And in 2021/22, virtual reality environments and Oculus headsets were recorded by police as being used in child sexual abuse image crimes for the first-time.[19]

Some of the risk factors for other types of service are relevant to VR, such as cross-platform risk and the manipulation of fake or anonymous profiles. However, there are also a number of risk factors which are specific to VR which should be addressed by Ofcom. For example, avatars are widely used, content and world creation is much more user-drive, and the nature of immersive environments, where senses are intensified, means harm can be experienced in very similar ways to the 'real world'.

Suicide and self-harm

We agree with the analysis in this section. It is particularly important to recognise that young people who are already struggling with their mental health are most likely to be at risk from this content. Evidence from Childline shows that young people who are exposed to suicide and self-harm content online find it particularly distressing if they are already experiencing mental health problems. We recognise this may be more relevant for the Children's Safety Code in the future, but it is critical that services understand that existing vulnerabilities can mean that content does not need to graphic or alarming to be harmful for these young people. Continued exposure to large quantities of content that idealises self-harm and suicide can and does have a significant impact on children.

Our recent evidence review found that the negative impacts of exposure to this type of content include triggering, distress, reinforcement and normalisation of 'pro' attitudes, suicidal ideation, as well as increased engagement in the related behaviours.[20] These effects may be further exacerbated where children are members of online communities that are supportive of these behaviours, as the social interaction they provide can further normalise and encourage them.

"I recently found self-harm content online, where you can watch people harming themselves or see pictures of it. I can't stop watching and searching for it. I used to self-harm, and this gives me the same feeling of triggering myself, but it makes me feel sick at the same time. I'm embarrassed I do it. I know I need to stop and don't know how. How else am I meant to cope?" Call to Childline from a girl, aged 13.[21]

[17] Camber, R. (2024) British police probe virtual rape in metaverse. Mail Online.

[18] Allen, C. and McIntosh, V. (2023) Child safeguarding and immersive technologies: an outline of the risks. London: NSPCC.

[19] NSPCC (2023) We're calling for effective action in the Online Safety Bill as child abuse image crimes reach record levels.

[20] Bryce, J. et al (2023) Evidence review on online risks to children. London: NSPCC.

[21] Please note that Childline snapshots are based on real Childline service users but are not necessarily direct quotes. All names and potentially identifying details have been changed to protect the identity of the child or young person involved. This applies to all snapshots uses in this response.

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

Governance and accountability

Question 3: Do you agree with our proposals in relation to governance and accountability measures in the illegal content Codes of Practice? Please provide underlying arguments and evidence of efficacy or risks to support your view.

Strong governance and accountability measures are critical to ensuring that risks are effectively identified, understood, and mitigated; we support these proposals as necessary measures to deliver these outcomes.

The proposal that all services should have a person accountable to the most senior governance body for compliance will help ensure regulatory compliance is consistently considered in decision-making at the very top of organisations. This requirement must, however, be distinct from the enforcement power which enables a senior manager to be held liable for compliance with a confirmation decision. The senior manager(s) held liable in these cases must be selected based on who has authority for the matter(s) covered in the confirmation notice – for example, content moderation, reporting, privacy settings. We discuss the implementation of senior manager liability further in Q53.

Question 4: Do you agree with the types of services that we propose the governance and accountability measures should apply to?

As noted above, we agree with these proposals. Volume 3 clearly outlines that these structures are necessary for consistent and robust risk mitigation and management – processes which are fundamental to the entire regulatory framework.

Some of the proposed governance and accountability measures should therefore be extended to apply to small services as well as large services, and, as a minimum, should apply to small services with specific or multiple risks. Without these processes, small services will be ill-equipped to systematically identify, manage and report on risk. Small services are certainly not immune to significant risk and there will be smaller services which grow rapidly. In these scenarios, it is vital that they have robust governance measures in place to ensure they are equipped to respond to changing risk profiles. We discuss risks on small services further in Question 14.

There are two measures in particular which should be applied to small services:

- Measures 3E: Tracking evidence of new and increasing illegal harm.
- Measure 3A: Annual review of risk management activities.

The measures set out in 3E are particularly important because services will not be able to adapt their risk mitigation strategies to growing threats if they do not have an up to date understanding of how their service is being used to facilitate illegal harms. Smaller services with specific/multiple risks are not exempt from most of the relevant measures in the Code of Practice, and so they must be able to meaningfully track new risks on their service to inform which measures they need to implement and ensure compliance. This measure is particularly important for ensuring all companies are proactive in tackling illegal harms, and do not wait until significant harm has already taken place and been reported before it is addressed.

Similarly, for 3A, services need to review their risk management activities to understand what is working and where they need to be strengthened. The online threat landscape is constantly shifting.

It is reasonable to expect that year-on-year, companies will need to adapt their risk management approach in response to this, especially if a service is at risk of illegal activity. The detail of this annual review would be proportionate to the size of the service, as it is likely that smaller services will have more streamlined governance systems, and so it is not an unreasonable burden to extend this measure for smaller services.

Effective governance will provide a bedrock for companies making decisions with user safety in mind. Without extending these measures to small (risky) services, the governance and accountability requirements on small services will be limited and consequentially risks undermining the whole of the regulatory regime.

Question 5: 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?

We urge Ofcom to reconsider the involvement of independent third-parties in monitoring and assurance for large multi-risk services.

There is a significant risk that companies will not put the strongest measures in place to tackle harm and protect users. Evidence from whistle-blowers has highlighted that tech companies are often unwilling to review internal data and tackle risk (discussed further in answer to Q7). Internal processes have clearly failed many large services to date, who have repeatedly prioritised profit and user engagement over children's safety. Moreover, during the passage of the Online Safety Act, some services were publicly hostile about the Act's aims and indicated they would be unwilling to comply with some parts of the legislation.[22] External review can help ensure that risk management systems fully meet the requirements of the regulation, identify where companies need to strengthen processes, and bolster Ofcom's supervision efforts.

External assurance can help improve transparency in regulatory regimes. In the water sector, Ofwat uses an information assurance regime whereby regulated companies provide information about their performance and compliance.[23] However, even within this established assurance regime, regulated companies have been found to deliberately misreport information in order to avoid penalties. This highlights the importance of having third party assured information and a record of activity.

Third-party auditing will be particularly valuable for services choosing to implement their own measures, rather than comply with the Codes of Practice. There will be significant challenges involved in evaluating whether these measures are reasonable equivalents of those recommended in the Codes. Third-party review will help reinforce the work that services and Ofcom will do to understand these approaches and reduce the risk that companies are able to overplay their efforts through providing an independent assessment of their efficacy.

As this recommendation is for large multi-risk services, it is reasonable to expect that they will have the resource to fund a third-party review. And given their mitigation processes will likely be the most complex (due to size and risk level), they also stand to benefit most from an external evaluation.

[22] Robison, K. (2023) Signal president says company will not comply with proposed U.K. 'mass surveillance' law requiring mandatory message scanning before encryption. Fortune.
[23] Ofwat. Information and assurance.

Question 6: 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?

We note that in other regulated sectors, albeit sectors where the regulator sets the revenue such as water[24], executive remuneration is tied to delivering positive outcomes. We would encourage Ofcom to consider how they can shift senior management's attention to safety outcomes.

Service's risk assessment

Question 7: Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views.

We are not best placed to comment on the specific risk assessment proposals. We are, however, concerned with the suggested definitions of core and enhanced evidence inputs for the risk assessments.

Internal evidence inputs

We are concerned that internal results and information which a service holds are not classed as a 'core input' for their risk assessment. This is illogical and is likely to weaken the accuracy of risk assessments. ***All internal information which a service has available should be considered in risk assessments.***

The results from product testing, content moderation systems, and assessments of previous interventions to reduce risk will be particularly important. Without using this data, services will be ill-equipped to effectively judge the efficacy of their current approach to risk mitigation and identify where their safety measures have not had the desired impact and further steps are required.

It is highly proportionate and reasonable to expect services which hold this data to use it. If services are not willing to use this data, it will signal a strong reluctance to genuinely engage with the risks they might pose. Accurate and comprehensive risk assessments are fundamental to this regulatory regime, and so every effort must be made to ensure this is a robust process.

External evidence inputs

As noted in response to Q5, some tech platforms have demonstrated an unwillingness to meaningfully engage with evidence of the risks on their services, to the detriment of the safety of their users. In 2021, whistle-blower Frances Haugen shared internal research by Facebook which had found that Instagram was negatively impacting the mental health of teenagers, including making girls feel worse about their bodies. Rather than acting on this research, Facebook buried it.[25] More recently, Arturo Béjar, previously an employee at Meta, argued the lack of transparency about the harms teenagers experience on Instagram meant they failed to base decisions in data about user's experiences and the safety settings on Instagram did not address the root causes of risk on the platform.[26]

Even with regulatory scrutiny, there is a real risk that services will seek to bury or underplay evidence of harm on their service. We do not believe the core inputs alone will provide a comprehensive overview of harm to inform robust risk assessments. Accurately identifying risk on a

[24] Ofwat (2023) Protecting customer interests on performance-related executive pay: 2022-23 assessment.
[25] Clayton, J. (2021) Frances Haugen: Facebook whistleblower reveals identity. BBC News.
[26] Kleinman, Z., Gerken, T. and McMahon, L. (2023) 'I blew the whistle on Meta, now I won't work again'. BBC News.

service is critical to the success of the regulation, as it informs if a service is low risk, specific-risk or multi-risk. The risk profiles are valuable, but ultimately only provide a starting point for a company identifying their key risk factors and their connection to illegal harm. In reality, especially for large services, there will be much more complexity involved in understanding how risk factors interplay with one another, and how the specific design and purpose of their service can enable illegal harms.

The risk of tech companies burying or underplaying evidence means independent experts must be able to feed into the risk assessments, to ensure services cannot underplay evidence of harm and have to act on it.

In particular, we recommend that the following enhanced inputs are instead categorised as core inputs for large services and for services with multiple risks or a specific CSEA risk.

- ***Views of independent experts [including NGOs]***
- ***Consultation with users and user research***
    - ***And / Or – Engaging with relevant representative groups***

Large and risky services are likely to face more complex harms, and they will need a strong evidence base to identify the full extent of illegal activity on their service, building on the risk profiles. Whilst we urge that these are included as core inputs for all services with a specific or multi risk, they should at least be required for those with an identified risk of CSEA. We welcome that there are separate decision frameworks provided for assessing the risks of CSAM and grooming, but for companies to meaningfully engage with these, external input and information from impacted groups will be vital.

Children and young people's voices and experiences must inform risk assessments. So too must survivors. If these groups are not involved, either through direct consultation or engaging with representative groups, risks assessments and mitigations may be detached from the reality of their experiences. Speaking with children and survivors will help identify how risks are changing on a service, the biggest risks they face, and any unintended consequences of design features or safety decisions.

For example, a platform might argue that developing an app which only encourages young people to add people they know in their local area, rather than creating connections across the world, is a safety mitigation to reduce the risk of grooming from unknown connections. But research with children by Revealing Reality found that the 'local' nature of Snapchat – which means that young people add lots of other young people in their area – creates a new type of risk.[27] The research found that the importance of local reputation to some vulnerable young people on the platform drove them to be involved, for example, in local fights which are then shared on the service. Young people felt pressure to turn up to these fights, either as a spectator or participant, because they knew this would be seen by young people across their local area via Snapchat.

Another case is Snapchat's 'My Eyes Only' folder. Snap have argued that this will enable users to keep images 'extra private' on their app. However, one young person working with the NSPCC previously raised that having a password protected folder indicates to young people that they can use this for images that they do not want shared with anyone else. This can lead to young people uploading images which put them at risk if shared. The young person knew of multiple young people (under 16) from her community who have had their accounts hacked and images from their 'My Eyes Only' folder shared publicly on Snapchat, including intimate images. They questioned why this

[27] Revealing Reality (2023) Anti-social Media – what some vulnerable children are seeing on Snapchat.

feature exists for under 18 accounts, considering the behaviour it encourages and the potential risks it exposes to young people using it.

It is by working directly with young people and child safety experts that these nuances and complexities are best understood. All services with a specific risk of CSEA, and certainly all large services, should therefore be required to have this input. For some, they will already have measures in place, such as advisory groups, which can help fulfil this role. This will balance the risk that companies will distort their own data through ensuring they are informed by accurate, independent assessments of the risk on their services.

Transparency

It is worth noting that for external organisations to effectively feed into a service's risk assessment process, greater transparency will be critical. At present, it is extremely challenging to understand the full scale of harm on a service, and what mitigations are available, because this information is not shared by platforms. Instead, civil society is often reliant on extensive research, undercover investigations, and the work of whistle-blowers. It is also essential for civil society organisations who design services which respond to harms to have transparent information from technology companies. We discuss this further in the information-gathering powers section, but Ofcom must consider how they will promote a culture of transparency in the new regulatory regime.

Robustness of risk assessments

A service's risk assessment will directly inform Code of Practice measures they implement. It is crucial that services are incentivised to carry out robust risk assessments which make every effort to identify the full range of harms on their platform, to ensure they have to comply with all relevant measures in the Codes.

We would welcome Ofcom setting out, in this Volume or in the enforcement guidance, that they will be working closely with regulated services, particularly supervised services, to ensure their initial risk assessments meet expected standards.

Setting out the intention now that Ofcom will be monitoring these first risk assessments emphasises the importance of this stage and will help ensure that the first set of actions implemented by services are appropriate and as strong as possible. This is particularly important because Category 1 services will only be required to publish a summary of their risk assessment, meaning Ofcom must provide the public and civil society with the confidence that regulated platforms will be delivering upon the ambitions of the Act.

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

Question 8: 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?

Question 9: Are the Risk Profiles sufficiently clear and do you think the information provided on risk factors will help you understand the risks on your service?

Record keeping and review guidance

Question 10: Do you have any comments on our draft record keeping and review guidance?

Question 11: 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?

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

Our approach to the Illegal content Codes of Practice

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

We support these first Codes of Practice as an important minimum standard, setting out a range of sensible measures to tackle illegal harm. However, many of the platforms which we are most concerned with would already be considered compliant when measured against these Codes and the benefits to children may only arise after the next iteration. As we highlight in our response below, we think Ofcom's proposed measures must be considerably more ambitious, particularly for risky services, in the future. Otherwise, the underpinning goals of the new regulatory regime to keep children safe online will not be met.

Engagement

We support the fast publication of this consultation following the passage of the Act. This will help ensure that there will be measures in place to hold services accountable for compliance as soon as possible. However, it is disappointing that there was not an opportunity for earlier engagement. In the development of future Codes, independent experts and civil society must be engaged much earlier in the process. This will be vital for ensuring the Codes are ambitious and well targeted before they reach consultation stage, where there is limited scope for making substantial amends.

Similarly, Ofcom *must* actively engage children and young people and survivors of online CSEA in the development of future Codes. It would be entirely unreasonable to expect these groups to respond to such detailed and lengthy policy documentation. Whilst some civil society organisations will work with these groups to inform their own responses, there must be formal mechanisms in place to ensure children and survivors are able to share their experiences and shape future Codes. They have the direct, first-hand insight necessary to understand how technological developments impact children, the risks they pose, and what works to keep them safe, which will be fundamental in developing more ambitious Codes going forwards.

Evidence and ambition

It is welcome that this approach has started with a thorough examination of the breadth of illegal harms children experience online and the incredibly harmful impact they can have (Volume 2).

However, we do not believe the scale of harm demonstrated in this assessment is reflected in the ambition of the Codes of Practice.

If companies are undertaking robust risk assessments using a comprehensive evidence-base, they are likely to uncover a wide-range of complex harms with multiple causes. But the structure of the Codes means, for the most part, to tackle these they will only need to implement generic measures which, in many cases, they will already have in place. This is particularly the case for large services. Our analysis of police data for online CSEA offences shows that in the majority of cases, it is the largest social media platforms (Snapchat or Meta-owned platforms Instagram, Facebook and WhatsApp) which are involved. Yet there will be few major changes that these services will be required to make when complying with the Codes.

We are concerned with the high evidential bar which has been set for proving the efficacy of the suggested Code of Practice measures. It cannot continue to be the case that only measures which are widely adopted within industry are recommended. This approach has a significant bias towards the interests of industry and will not bring the systematic changes required to better protect children online.

This approach could also remove the incentive for platforms to invest in and rollout ground-breaking safety measures. Currently, many services are going beyond the measures proposed in the Codes. However, if platforms are deemed as compliant by only implementing the Code measures, it will be difficult for Trust and Safety teams to justify investing in new solutions. Internal decision makers may favour rolling out older technology recommended in the Codes over new, innovative measures, regardless of how impactful, because they are not necessary to be deemed compliant and in fact risks setting a higher bar for themselves. If all platforms take this approach, the Act will fail to achieve its objectives.

In some cases, it may be necessary for Ofcom to set out results or outcomes for services to achieve, such as reducing the number of fake profiles created, without recommending specific measures to do this. Ofcom are well placed to identify which risks need tackling, but in some cases, it will be services who have the expertise and resource to develop solutions which are appropriate for their platform. This should also provide Ofcom with more flexibility when making recommendations.

In future Codes we would welcome Ofcom setting outcomes that services should meet to help reduce illegal offences on their service; it should not always be necessary to include the means for achieving this.

We would welcome a further discussion with Ofcom on the appropriate evidential threshold to recommend measures within Codes. This is a key issue that needs to be resolved if, going forward, Ofcom's Codes are going to be effective in addressing the harms and risks that children and young people are experiencing online.

<u>Safety by Design</u>

Many of the measures proposed by Ofcom in the Codes of Practice focus on mitigating harm after the fact, for example by moderation once a piece of content has been shared and user reporting. They ultimately recognise that there will be some level of harm on a service and look to how these can be mitigated.

Moving forwards, as well as these measures, we would like to see a greater focus on challenging services to build platforms which are safer in their design. This should include identifying where greater friction can be introduced in offending pathways (discussed in Q16) and ensuring recommender systems are not promoting material which is illegal / facilitates illegal activity.

Question 13: 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?

We recognise the need to consider proportionality when recommending measures and the logic of focusing on large and/or risky services. However, there are limitations to this approach. As recognised by Ofcom in Volume 2, evidence indicates that services which are prioritising growth may do so at the expense of safety measures, which can be exploited by CSEA perpetrators. This means that small services seeking to grow rapidly may initially overlook safety, and the Codes of Practice will do little to rectify this if they are not initially identified as risky.

Our response to Volume 3 and answers below highlight areas where measures should be extended to small services as well, to ensure that they are embedding effective safety measures and governance systems which can be adapted and built on if and when they expand. Our recommendations would not result in a disproportionate burden on these services but ensure there is a minimum level of safety activity in place which it is reasonable to expect any platform to consider.

Question 14: Do you agree with our definition of large services?

Seven million users sets a high bar for a large service. Once the full regulatory regime is in operation, we would expect that measures which are currently only expected for larger platforms to also apply to smaller ones. As a first step, this should mean applying key measures to risky services with over 700,000 users.

We raise two further issues for consideration and clarification below.

Large services for children

A large service is defined as a service with a number of monthly UK users that exceeds 7 million – roughly 10% of the UK population. This proposal overlooks services with a low adult but high child user base. There are just over 14 million children in the UK.[28] If a service is used by around 10% of this population (1.4 million), this would make it a large service for children, but it would fall well below the proposed definition. All Volumes consistently suggest that applying measures to services with the highest reach is likely to have the greatest impact for user safety. However, this risks excluding services with a high concentration of users from vulnerable or marginalised populations – including children. It is vital that those services which are most popular amongst children have to identify and mitigate risks (particularly of CSEA).

Limited data availability means it is difficult to know how large many platforms are, but two platforms which may not currently be classed as large, but would be large for children, are Kik and Yubo. Evidence suggests that both these platforms are likely to have less than an average of 7 million monthly users in the UK.[29] However, they are particularly popular apps amongst young people and so would likely be a large service for children. They also pose a risk to children's safety. Police data shows that Kik was involved in over 450 grooming cases recorded by UK police from 2017/18 to 2022/23.[30] It would therefore significantly benefit children's safety if some of the measures which do not apply to smaller services (e.g. on governance and accountability) were extended to these platforms.

We recommend that Ofcom develops a new category of 'large services for children' in future Codes and applies key child safety measures to these platforms.

Growth

Clarity should also be provided regarding how often services need to review their user-base to determine when they have become a large service. Volume 4 encourages service to 'keep track of how their average user figure fluctuates month by month' but does not provide any specific requirements. Clear expectations for tracking growth are important for ensuring that companies comply as large services once they meet the 7 million threshold.

[28] Unicef (2023) How many children are there in the UK?
[29] Woodward, M. (2023) Kik 2022 User Statistics: How many people use Kik? Search Logistics; ParentZone (2023) Yubo.
[30] NSPCC analysis of grooming data from UK police based on FOI requests from 2017/18 to 2022/23.

We recommend that companies are required to track their user base on a monthly or quarterly basis.

Ofcom should also track the impact of the decision to use a 12-month average for measuring userbase. There are circumstances where this could mean that a service rapidly grows and reaches 7 million users, but then falls below this threshold again within 12 months. Whilst this may be unlikely, it is important that all large services are held accountable for tackling illegal harms. Ofcom must ensure that this metric is effectively capturing all services with a significant reach and impact.

Question 15: Do you agree with our definition of multi-risk services?

Yes, we agree with the definition of multi-risk services applying to services with at least two different kinds of illegal harm. More than one illegal harm on a platform indicates that it is used by different types of bad actors, requiring a more complex and comprehensive response.

Question 16: Do you have any comments on the draft Codes of Practice themselves?

There are a number of other areas that should be addressed in the Codes of Practice. We highlight some of these in our answers below. Others do not align directly with the categories in the Codes and so we have provided further detail in this answer.

Private messaging

One of our primary concerns is the lack of requirements for private and end-to-end encrypted (E2EE) services. We recognise that there are limitations imposed by the Act. However, the proposed measures and exemptions mean that the expectations on these services are severely limited, and for large platforms such as WhatsApp they will not need to introduce any substantial changes.

Private messaging is the frontline of online grooming. Data from ONS shows that 74% of approaches to children by someone they do not know online first take place via private messaging.[31] These platforms are exploited by offenders, who move children that they have met on more public platforms to these sites for exploitation. This is because of the well-known challenges of detection CSEA in E2EE platforms. The NCA have highlighted that the roll out of E2EE on Facebook Messenger and Instagram could mean that the alerts they receive from Meta via NCMEC, which enable them to find perpetrators and safeguard children, could fall by 92%.[32]

"Basically I am being blackmailed by someone I met on Wizz, an app for meeting new friends. **This person, who claimed to be a 16-year-old girl from America, told me to go on WhatsApp and send her explicit pics of myself.** Now she's threatening to post the pics to my friends on social media unless I pay her £50. She thinks I'm on my way to getting the money but I'm not, I don't have that much. I'm really scared." Call to Childline from a boy, aged 17.

"I'm in a serious situation that I want to get out of. **I've been chatting with this guy online who's like twice my age. This all started on Instagram but lately all our chats have been on WhatsApp.** He seemed really nice to begin with, but then he started making me do these things to 'prove my trust' to him, like doing video chats with my chest exposed. Every time I did these things for him, he would ask for more, and I felt like it was too late to back out. I feel so stupid for even going this far. I thought I knew how to keep myself safe but clearly I don't. This whole thing has been slowly destroying me and I've been having thoughts of hurting myself." Call to Childline from a girl, aged 15.

[31] Office for National Statistics (2021) Children's online behaviour in England and Wales: year ending March 2020.
[32] Symonds, T. (2023) Facebook encryption risks children's safety, National Crime Agency warns. BBC News.

Tackling online CSEA demands action from private and encrypted services. ***As far as the limits of the Act allows, future Codes must introduce stronger requirements for private messaging.*** This is discussed further in answer to Q31.

Ofcom should also introduce non-binding guidance for private messaging and end-to-end encrypted service which provides detail on how and when Ofcom will look to use proactive technology notices to deal with CSEA content.

Detecting suspicious patterns of activity

Large and risky services should have tools in place to detect suspicious patterns of activity on their platform. For example, on Facebook and Instagram Meta uses machine learning tools to detect patterns of abuse, identifying signals such as users adding accounts with content sexualising children, using coded language in posts and bios, and searching for egregious terms.[33] Meta have been able to monitor these individuals and disrupt their networks, disabling accounts that violate their child safety policies. WhatsApp's detection efforts include using automated technology to proactively scan unencrypted information for suspected CSAM sharing, such as by analysing individual and group profile photos, group descriptions, and group behaviour.[34] Whilst this is not as effective as monitoring content data, it shows that there are possible upstream mitigations for all service types.

This practice should be recommended in future Codes to ensure that services are proactively identifying dangerous actors, and not relying on children to report illegal activity. This will be particularly important for tackling perpetrator networks who share CSAM, where reporting tools will be futile.

Cross-platform risk

As recognised in Volume 2, abuse is often not siloed to one platform. One common grooming tactic is for offenders to redirect conversations from public spaces to more private channels, including end-to-end encrypted environments. Abusers also speak to children simultaneously on platforms, such as by meeting and interacting with a child on a gaming platform, whilst actively grooming them on an ancillary chat platform. Having multiple accounts across different platforms enables offenders to evade attempts by children to block them or stop engaging with them.

"We met around the start of quarantine on an online video game. It was purely by coincidence that we ended up on the same team. We did quite well so we decided to party up for another round. There is an in-game voice chat function for those in a party, so we ended up being friends over the next few games. We had a lot of conversations via chat, and after a few months we transitioned to Discord then eventually WhatsApp. We now talk every day and have made plans to meet in person. I need advice about it from others as most people do when dating someone, but no one seems to be able to look past his age." Call to Childline.

"I met this guy on Discord and we became friends. After a while he started buying me virtual items in games. At first I thought it was just a friendly thing but then he told me he liked me. When I told him I didn't feel the same way he got upset and started acting real weird, like asking me for nudes. I blocked him after that, but then he somehow found my other social media accounts and has been

[33] Meta (2023) Meta's Approach to Safer Private Messaging on Messenger and Instagram Direct Messaging.
[34] WhatsApp Help Centre. How WhatsApp Helps Fight Child Exploitation.

posting lies and mean comments about me ever since. I wish I'd never met this person, he's making my life hell." Call to Childline from a girl, aged 13.

During the passage of the Online Safety Act, Government made explicitly clear that in order to achieve their illegal duties, including tackling the facilitation of CSEA, they expect services to consider cross-platform risk.[35] In Committee Stage, the Government argued that, as part of their illegal content risk assessment duty to prevent users from encountering illegal content 'by means of a service', services must consider instances where users are directed onward to illegal content on another site. The Government also raised that services must collaborate reasonably to address illegal harms which span multiple platforms. In both instances, they noted that Ofcom could support services to deliver this through the Codes and enforcement action.

There are currently limited examples of how services can tackle cross-platform risk. As part of Roblox's chat filtering system, they block users from sharing personal information and from receiving instructions about how to move off the platform.[36] The announcement of Lantern, a cross-platform signal sharing programme to tackle grooming and financial sextortion of young people, appears to be a leading example of how services can collaborate and share information to prevent abuse.[37]

In future Codes, Ofcom should assess existing systems and address how services can identify perpetrators and limit the risk of children being groomed across multiple services.

Targeting perpetrator behaviour

It is vital that further measures are added in the next iteration of Codes which directly target perpetrator activity.

A common feature of online grooming is perpetrators creating fake profiles to connect with children. In particular, there is significant evidence from Childline showing that offenders create fake profiles so that they can pretend to be children and convince the children they are talking to that they are speaking to someone of a similar age. These interactions often lead to children sharing images or personal information with these fake profiles, which is then used to blackmail them to send nude images or money.

"I really thought this guy I was talking to was my age. We had been texting for a while, then it became sexting and then he asked for nudes. I sent some but he insisted I send more with my face in. He just kept asking and peer pressuring me so I just did it. Now he's told me he's 32! I don't want to talk to him anymore, but he has my pictures and is threatening to share them. They're clearly of me! I'll take a look at the Report Remove page on the Childline website and think about if I want to report this" Call to Childline from a girl, aged 16.

"I was talking to this girl on Snapchat who I was thought was my age, then she said she was actually much older and got angry I didn't want to speak to her anymore. She made fake sexual pictures of me and demanded I send her £200, or she'll send it to my friends. I've reported and blocked the account but don't know how to be sure they won't send the pictures." Call to Childline from a boy, aged 16.

To reduce the risk of children be targeted by these profiles, services must tackle the creation of fake profiles – and particularly profiles where adults are claiming to be children. When recommending age verification and age assurance methods, Ofcom should consider not only the

[35] House of Commons Official Report – Public Bill Committee – Online Safety Bill (2022).
[36] Roblox. Safety & Civility at Roblox.
[37] Tech Coalition (2023) Announcing Lantern: The First Child Safety Cross-Platform Signal Sharing Program.

importance of accurately identifying children, but also the importance of identifying and blocking spurious accounts.

Future Codes must also target the facilitation of abuse. For example, perpetrators collate and share large collections of legal imagery of children (such as children in swimwear). Individually, these items are entirely innocuous and very unlikely to be flagged to moderation individually.[38] However, services can and should identify users who are collating these images and sharing them with others at scale to detect and block bad actors.

[The rest of this answer is confidential]

The COM-B CSA Behavioural Analysis Project, carried out by the NCA, provides analysis of offender behaviour and outlines the main capabilities that facilitate grooming.[39] There are a number of areas where proactive tools by tech companies could identify and disrupt perpetrator behaviours, including at these stages:

- Detecting the downloading of multiple messaging apps over a short period of time to identify risky accounts on, for example, app stores.
- Identifying concerning search terms, such as "how to access dark web", to detect risky accounts.
- Creating friction in downloading of multiple images that might be used to create fake profiles.
- Detecting and flagging membership of groups where child sexual abuse is normalised or acceptable.
- Identifying adult accounts with lots of contacts to children's accounts to assess for potentially risky activity.

We urge that future Codes contain measures which directly target perpetrators in order to prevent abuse, considering existing research about perpetrator behaviour online. We would welcome the opportunity to discuss how abuse is facilitated and could be tackled in future Codes with Ofcom in more detail.

Question 17: Do you have any comments on the costs assumptions set out in Annex 14, which we used for calculating the costs of various measures?

Content moderation (User to User)

Question 18: Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views.

Effective content moderation is an important safety tool. Notably, it is largely reactive, so it is critical that systems are as accurate and efficient as possible to limit the spread of illegal content if it is shared. Services are likely to use both human moderation and automated tools (such as AI) for moderation. Automated systems have a critical role to play, particularly for rapidly identifying illegal material and for effective moderating vast amounts of content on large services. Systems should be supported by appropriate human input, including to ensure that the most egregious illegal material is removed as an urgent priority.

[38] Crisp: A Kroll Business (2023) COITP: How predators hide in plain sight.
[39] Confidential.

Online services should be expected to quality assure their moderation systems to ensure that they deliver the correct outcome and that they dial up or down the role that human and automated moderation plays as appropriate.

We set out comments further for three of the proposed measures below. We support the approach taken with the other measures.

4B: Setting internal content policies; and 4F. Provision of training and materials to moderators.

We support these proposals, however ***we urge that they are also applied to small services with a specific risk***. Establishing effective content moderation systems which are accurate and timely requires clear internal policies and up to date training and support for moderators. Both these elements will be essential for effective content moderation and decision-making on small services tackling a specific illegal harm.

Only having one illegal harm on a service does not mean that it is simple to moderate. For example, the Illegal Content Judgement Guidance recognises that there are a number of complexities in identifying and responding to grooming. Moderating grooming is recognised as particularly challenging due to its gradual nature (in some instances), the importance of context in identifying grooming patterns, and the necessity of human moderators.[40] This challenge will only be exacerbated by emerging technologies; for example, there is already evidence of child sexual exploitation in virtual environment multi-user spaces, such as in VR strip clubs, where children can be manipulated into performing dances for viewers in exchange for money.[41]

Without a robust content moderation policy and targeted training, moderators are likely to face significant challenges in identifying this harm, with inaccuracies or delays in decision-making posing a major risk to children's safety on and offline. It is entirely proportionate to expect that any service that has a medium or high risk of an illegal harm implements the systems and processes required to stand up an effective content moderation function that is targeted to the nature of harm on its service.

A4.17: The provider should resource its content moderation function so as to give effect to its internal content policies and performance targets

We strongly agree with the importance of this recommendation. Services with a risk of CSAM or grooming must resource their content moderation function to a standard which ensures they can consistently address these harms.

The dangerous impact of significant cuts to content moderation team has recently been demonstrated by X (formerly Twitter). Insiders from X have raised that substantial cuts to the content moderation teams have directly impacted the service's ability to tackle CSEA.[42] One previous employee noted that the team had already been understaffed, but the cuts resulted in a significant loss in expertise that meant that their capacity to proactively identify accounts sharing CSAM would be severely restricted.[43]

This measure should therefore be expanded to require services to have regard for the results of their risk assessment when resourcing their content moderation functions. This will ensure services have the appropriate expertise within their moderation teams to deal with the most pertinent risks

[40] INHOPE (2022) The importance of human content moderators.
[41] Allen, C. and McIntosh, V. (2023) Child safeguarding and immersive technologies: an outline of the risks. London: NSPCC.
[42] Spring, M. (2023) Twitter insiders: We can't protect users from trolling under Musk. BBC News.
[43] Spring, M. (2023) Twitter insiders: We can't protect users from trolling under Musk. BBC News.

to their platforms. For example, services with a significant risk of CSEA must have individuals within their content moderation teams with the expertise and experience to moderate this harm.

We welcome the recognition that services should also consider the languages used by their user-base. Whilst research on this issue is limited, there is evidence to suggest that moderation efforts are substantially weaker for languages outside of English. For non-illegal content, the impact of this is particularly clear for disinformation – again, cuts to X's moderation teams means they have limited staff who understand local language and cultural references, which has resulted in significant challenges tackling disinformation.[44]

Search moderation (Search)

Question 19: Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views.

There are two main risks to children posed by search services. Firstly, there is a risk that children can access dangerous illegal material, such as illegal self-harm and suicide content, through search engines. Secondly, perpetrators can search for and access CSAM through search engines.

We therefore support these proposals as well-targeted and proportionate measures, which will make it more unlikely that users are able to actively find or accidentally stumble across illegal activity and content on search engines.

Automated content moderation (User to User)

Question 20: Do you agree with our proposals? Do you have any views on our three proposals, i.e. CSAM hash matching, CSAM URL detection and fraud keyword detection? Please provide the underlying arguments and evidence that support your views.

As noted in answer to Question 18, automated content moderation tools are critical for identifying illegal material at pace. Future Codes must give much greater weight to the important of these technologies – either by recommending services implement specific tools or by requiring them to develop their own. This approach must be supported by human input, to ensure these systems are accurate and that legitimate reports are not overlooked by automated tools.

Hash matching and URL detection

We strongly support the proposals for CSAM hash matching and CSAM URL detection. These tools are critical for ensuring CSAM can be proactively detected, removed, and reported, avoiding a reliance on user reporting and preventing the spread of these images in the first place.

Taking a proactive, preventative approach to the distribution of CSAM is fundamental to protecting the rights of these victims. Ensuring that the number of people who see this material is limited as much as possible is critical for protecting the privacy of the victim. We also know that the knowledge that other people might view these images, or use the images to find them, can negatively impact children's mental health, including by making them angry, distressed, or scared, and mean the child feels re-victimised.

CSAM URL detection is particularly important for tackling the way perpetrators use social media to network and direct other perpetrators to sites hosting CSAM. As well as forming offender networks

[44] Shih, G., Miller, M. and Menn, J (2023) Twitter hate speech up in large foreign markets after Musk takeover. Washington Post.

online, we know that perpetrators 'sell' access to CSAM URLs.[45] In an effort to evade detection, they may not refer explicitly to CSAM, but instead use coded language and other content (e.g. legal images of children) which is recognisable to other offenders. Detecting URLs as well as CSAM is therefore critical for disrupting this activity.

The important impact on children's rights of these tools can be contrasted with the negligible risk to user privacy for non-victims. The National Centre for Missing and Exploited Children (NCMEC) estimate that the hash matching tool PhotoDNA has a false positive rate of under one in one trillion.[46] Combined with human moderation, the likelihood of an individual being misidentified for sharing known CSAM is incredibly low.

Currently, CSAM hash matching tools are used by many large platforms where content is not end-to-end encrypted.[47] We urge Ofcom to monitor the availability of these tools and expand the types of service this measure applies to. ***As a priority, this measure should also be applied to smaller services (700,000 users) with a medium-risk of CSAM (rather than just a high-risk as is currently required).***

Detecting 'new' CSAM

It is a major gap that this Code of Practice does not contain any measures to support the proactive detection of unknown or 'new' CSAM.

Known CSAM represents just a fraction on online CSAM, but due its nature it is incredibly challenging to determine the scale of unknown material. Based on data from NCMEC, it is estimated that every second at least two images or videos of child sexual abuse are published online.[48] Within six months in 2021, services using Google's machine-learning tool to proactively identify 'new' CSAM classified over six billion images.[49] While the true scale cannot be understood without the widespread implementation of proactive technology, it is clear that a huge number of children who are victims of CSA will be overlooked under the current measures. It is unacceptable that all victims of child sexual abuse material will not receive the same protections, given the severity of this harm.

The importance of platforms detecting new CSAM is reinforced by the challenges posed by the rapid development of generative-AI. This technology is already being used at scale to generate hyper-realistic CSAM.[50] If proactive technologies are not in place to detect and remove this content, there is a risk that masses of synthetic CSAM, indistinguishable from non-AI material and based on images of real children, will become widely available and overwhelm moderation teams.

Industry approaches to detecting new CSAM are much more varied than for known CSAM.[51] Many platforms are not using specific tools for this, but there are examples of good practice. Google uses artificial intelligence to identify new material.[52] Their system is designed to limit the risk of false

[45] This evidence has previously been shared with NSPCC from journalists undertaking investigatory work.
[46] Steinebach, M. (2023) An Analysis of PhotoDNA. In *The 18th International Conference on Availability, Reliability and Security (ARES 2023), August 29--September 01, 2023, Benevento, Italy.* ACM, New York, NY, USA 8 Pages. https://doi.org/10.1145/3600160.3605048.
[47] eSafety Commissioner Australia (2022) Basic Online Safety Expectations: Summary of industry responses to the first mandatory transparency notices.
[48] Eurochild (2023) Last call: protect children from online sexual exploitation!; NCMEC (2023) CyberTipline 2022 Report: CyberTipline Files.
[49] Google. Fighting child sexual abuse online: Content Safety API.
[50] Internet Watch Foundation (2023) How AI is being abused to create child sexual abuse imagery.
[51] eSafety Commissioner (2022) Basic Online Safety Expectations: Summary of industry responses to the first mandatory transparency notices.
[52] Jasper, S. (2022) How we detect, remove and report child sexual abuse material. Google: The Keyword.

positives, and is reinforced by a specialist team of moderators who review imagery that it is flagged to ensure it is CSAM before it is reported, significantly minimising risk to user privacy. Meta also uses Google's Content Safety API in combination with an internal CSAM classifier to identify new images on Facebook and Instagram where end-to-end encryption is not enabled.[53]

There are a number of other tools which have been developed to enable companies to proactively detect new CSAM. In partnership with the Internet Watch Foundation, SafeToNet have developed SafeToWatch – a technology that can detect both categorised and uncategorised CSAM using machine learning.[54] The Vigil AI CAID Classifier uses the Child Abuse Image Database to detect and classify new image and video CSAM.[55]

The rights implications of CSAM (known or unknown) being shared online have been discussed above and must be given significant weight in decisions about recommending further proactive technology, given the severity of this harm and its priority status within the Act. Child sexual abuse online cannot be comprehensively tackled if companies are not working to identify and remove all forms of CSAM. As noted with Google's system, reinforcing proactive technologies with well-resourced and highly-trained human moderators can ensure that the risk of false positives being reported is significantly limited.[56]

The next Code of Practice must require both large and risky services to deploy tools which automatically detect and remove all forms of CSAM. Without tackling both known and unknown material, companies will not be able to meet their duties to prevent individuals from encountering priority illegal content and protect children.

Livestreaming

Detecting and disrupting CSEA on livestreaming is another area where automated technologies have a critical role to play. Livestreaming enables abusers to control and coerce children into increasingly extreme forms of abuse which can last for a long period of time and have a devastating impact on children.[57]

"I was so terrified and didn't know what to do. I was just online **and a friend of a friend asked me for a video call and to show nudes**. I did it, but **he'd recorded it and shared it on social media**. I wanted to tell my parents, but I don't want them to ban me from social media for one mistake." Call to Childline from a boy, aged 14.

"I'm slowly coming to terms with something that happened a couple of years ago. Basically I was groomed online, even though I didn't know it at the time. **This guy - who was 10 years older than me - made me do skype video calls and it was always something sexual**, like talking dirty or showing off parts of my body. He said he loved me but now I realise that was just his way of controlling me." Call to Childline from a girl, aged 17.

The NSPCC works with one family who have campaigned anonymously to raise awareness of the current risks of the online world to children and call for much stronger protections. Their experience illustrates the severity of CSEA committed through livestreaming:

---

[53] eSafety Commissioner (2022) Basic Online Safety Expectations: Summary of industry responses to the first mandatory transparency notices.
[54] SafeToNet, Detecting and preventing Harmful and Illegal Content in Images and Videos.
[55] Roke, Vigil AI.
[56] Jasper, S. (2022) How we detect, remove and report child sexual abuse material. Google: The Keyword.
[57] We Protect (2018) Global Threat Assessment: Working together to end the sexual exploitation of children online.

Ben* was 14 when he was tricked on Facebook into thinking he was speaking to a female friend of a friend. This person turned out to be a man. Using threats and blackmail, he coerced Ben into sending abuse images and performing sex acts live on Skype. The images and videos were shared with five other men who then bombarded Ben with further demands through livestreaming websites and other platforms. His mum, Rachel*, said: "The abuse Ben suffered had a devastating impact on our family. It lasted two long years, leaving him suicidal".[58]

The IWF have raised that the material they remove online is often taken from livestreams where children have been groomed, coerced, and blackmailed into streaming their own sexual abuse online.[59] Images and videos from these livestreams are clipped and then widely shared and sold by perpetrators. Preventing CSEA on livestreaming will be critical to protecting children and disrupting the trade of CSAM between perpetrators.

Industry efforts are currently limited, but there are cases where technology is effectively being used to protect children in livestreaming. On Yubo, algorithms and human moderation are used in combination to detect and interrupt CSEA in real-time, enabling Yubo to disable livestreams and suspend accounts when this activity is identified.[60] Alternatively, some platforms scan for potentially inappropriate comments posted alongside livestreams to detect CSEA.[61]

Given its high-risk nature, we would expect a future CSEA Code of Practice to set out measures requiring companies to tackle CSEA in livestreaming.

Keyword detection

Automated content moderation technologies should also be implemented to detect and disrupt grooming. The recommended measures in the Code of Practice for grooming rely on changing settings on children's accounts. Whilst these are important and necessary changes, they continue to place the burden on children to have to consider how they want to manage their safety online, and ultimately to be able to identify and report bad actors. They also focus, importantly, on preventing grooming – but where perpetrators are able to connect with children, these measures will then do little to disrupt the grooming process.

Online grooming is often detected rather than disclosed. In many cases, children can be the last ones to recognise that they are victims of abuse. The very nature of grooming, which often relies on deception, forming 'friendships' and building false trust, creates significant barriers to children being able to distinguish between benign and dangerous contacts. For too long the burden has been on children to recognise and report their own abuse, and as a result children have been unsupported and online grooming levels have been able to reach record levels. This approach must be changed. Machine learning alone will not be able to detect grooming, but it can help to significantly speed up the process and inform triaging and human moderation.

Keyword detection and machine learning offer technological solutions to improve grooming moderation. For example, at Swansea University, Project DRAGON-S is currently working with law enforcement to trial a new machine learning tool that will help them to quickly analyse chat logs to detect high-risk interactions.[62] This information is then being used to prioritise human moderation of chat logs to identify children at greatest risk.

---

[58] Names have been changed to protect anonymity.
[59] IWF (2018) IWF research on child sex abuse live-streaming reveals 98% of victims are 13 or under.
[60] Yubo, Safety Hub: Safety Tools; Yubo, Real-Time Intervention on Social Video.
[61] IICSA (2022) The Report of the Independent Inquiry into Child Sexual Abuse: F.3: Detecting online child sexual abuse.
[62] Project Dragon-S (2023) Project Dragon-S - Developing Resistance Against Grooming Online. Swansea University.

Another example is Roblox, which uses filtering technologies to assess all text chat on their platform to block inappropriate content, including sexual content, personal information, and instructions on how to move off the platform. These filters are updated on an ongoing basis. Implemented effectively, blocking this chat could directly target key points of risk in the grooming pathway.[63]

In future CSEA codes, services should be required to utilise a range of tools, including default settings, human moderation, and automated tools, to both prevent and disrupt grooming.

Question 21: Do you have any comments on the draft guidance set out in Annex 9 regarding whether content is communicated 'publicly' or 'privately'?

Strengthening the guidance

Annex 9 recognises the challenges of determining whether content has been communicated publicly or privately. It will often be the case that platforms have both public and private elements, and it will be vital that services accurately identify these to ensure that the strongest Code measures are implemented wherever possible. Whilst this guidance is a helpful start in making this distinction, significant further clarity is required.

Of the three key factors which will be used to determine if content is public/private[64], measure A – 'the number of UK individuals able to access the content', is particularly opaque. No definition is provided as to how many individuals content should be accessible to in order to be classified as public. As a result, there will be a significant level subjectivity in this process, and it is unlikely that providers will use a consistent definition of a 'substantial section of the public'. ***In the guidance, a specific threshold or range for what constitutes a 'substantial section of the public' should be provided.***

Given the complexity of this guidance, a number of case studies and examples should be provided to help clarify more complex issues. This is a valuable aspect of the ICJG which would also work well here.

For example, on Snapchat, group chats can be created with up to 1,001 users who do not all need to be connected to each other. Whilst this may not meet the threshold of a 'substantial section of the public', it is also unlikely that anyone using a group chat of that size, with so many users they are not connected to, would reasonably view their communications as private.

Another example is large WhatsApp groups. Whilst we recognise that many aspects of WhatsApp indicate that it is private communication, group chats can be widely accessible, with users able to find and join group chats with hundreds of users through clicking on publicly accessible links which anyone could access and use. Group chat membership means users can belong to a group with up to 1,024 participants. The 'communities' feature allows users to join groups of 1,000s by clicking an online advert for the community; many others in the community will not be known to the user.

"A while ago I saw a video on YouTube about how a guy was busting paedophiles and creeps on the internet by pretending to be a kid, and I kind of wanted to do a similar thing. I looked around Instagram for the creepiest accounts about kids my age and younger. In the end, I came across this link on one of their stories. It's a link to a WhatsApp group chat in which [child sexual abuse material] is sent daily! There are literally hundreds of members in this group chat and they're always calling the kids 'hot' and just being disgusting." Call to Childline from a boy, aged 15.

[63] Roblox, Safety & Civility at Roblox.

[64] (A) Number of UK individuals able to access the content, (B) Access restrictions, (C) Sharing / forwarding of content.

It would therefore be beneficial to include examples of a number of ways that content can be communicated and outline whether Ofcom would classify them as private or public based on the three criteria.

We suggest that Ofcom gives further consideration as to how regulated services are likely to understand this guidance to avoid the potential for misinterpretation, which would result in some parts of the Codes not being applied.

Metadata

We are concerned that this guidance [Annex 9] classes metadata as private content. Metadata plays a key role in services' efforts to detect and disrupt bad actors on their platforms. This is particularly necessary on encrypted services. Meta have already outlined that they will use metadata alongside other publicly available information to identify potentially malicious actors on their services as they roll out end-to-end encryption.[65] WhatsApp also collect and share metadata internally.[66]

Considering these platforms use or will continue to use metadata whilst end-to-end encrypted, we urge Ofcom to reconsider designating this content as private. This risks significantly undermining efforts to tackle CSEA on private messaging, and will limit the steps that private communication platforms can be asked to take to tackle illegal activity on their platform.

Encryption

Annex 9 only makes the distinction between content communicated publicly and privately. It does not set out different standards of private communication – for example, the difference between non-encrypted, encrypted, and end-to-end encrypted messaging services. Whilst we support this, we note that it is inconsistent with Ofcom's approach in other documents. We question why in the *Summary* and *Consultation at a Glance*, end-to-end encrypted services are specifically referenced as being exempt from certain measures alongside private communications.

The Act makes no specific reference to end-to-end encrypted services, it only refers to public or private communications. This technologically neutral approach must be maintained in the implementation of the regulation. Whilst in these examples end-to-end encryption is referenced alongside private communications, it is vital that end-to-end encryption is not given specific carve outs or special status within the Codes and wider regulatory regime. The Act's tech neutral approach is fundamental to future proofing and ensuring that there are no incentives for services to encrypt in order to evade regulatory responsibilities. We urge that Ofcom removes the specific references to end-to-end encryption and instead consistently uses the public / private distinction.

Do you have any relevant evidence on:

Question 22: the accuracy of perceptual hash matching and the costs of applying CSAM hash matching to smaller services;

Question 23: the 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;

Question 24: the costs of applying our CSAM URL detection measure to smaller services, and the effectiveness of fuzzy matching for CSAM URL detection;

---

[65] Meta (2023) Meta's Approach to Safer Private Messaging on Messenger and Instagram Direct Messaging.
[66] O'Flaherty, K. (2021) All the data WhatsApp and Instagram send to Facebook. Wired.

Question 25: the costs of applying our articles for use in frauds (standard keyword detection) measure, including for smaller services; and

Question 26: 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.

Automated search moderation (Search)

Question 27: Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views.

User reporting and complaints (U2U and search)

Question 28: Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views.

We support the aim of these measures to significantly strengthen platform reporting tools. At present, reporting is under-used and ineffective. Children are disillusioned with the efficacy of reporting tools; they think that reporting will not change anything, which disincentivises them from taking action.[67]

We caveat our response to this question with a note that reporting tools have significant limitations as a child protection measure. Children are often unaware of the dynamics of being groomed which stops reporting and means the most significant online harms are unlikely to be captured by reporting or complaints mechanisms. Offender-to-offender interactions will also not be tackled by reporting, as offenders who are sharing, for example, CSAM are highly unlikely to report someone else. Whilst we welcome efforts to improve reporting, this must always be balanced with a wider range of measures that ensure abuse is proactively detected and disrupted.

<u>New measure: Information about the reporting process</u>

We recommend an additional reporting measure is introduced: All services should be required to provide clear and accessible information about what the reporting process is, and what the consequences will be for the reporter and the user/content being reported, at the start of the process.

Reporting is often viewed by children as a 'black box'. Insights from Childline show that children can be put off reporting because they do not know what the outcome will be, and are worried that they will be negatively impacted. This can include worries about their anonymity, or that they have done something wrong by accidentally viewing illegal behaviour. This is reinforced by research from Thorn, which has found that one of the top reasons children do not report is because they are worried about remaining anonymous.[68]

"I only feel comfortable telling you this because of the confidentiality promise. I'll admit I watch a lot of porn, but yesterday I accidentally stumbled onto some [child sexual abuse material]. I immediately closed the website in a panic. I'm really worried about getting in trouble for even looking at it, even by accident. Thank you for explaining I can report it to IWF without getting into trouble, I'm going to do that." Call to Childline from a boy, aged 14.

[67] Thorn (2021) Responding to Online Threats: Minors' Perspectives on Disclosing, Reporting, and Blocking.
[68] Thorn (2021) Responding to Online Threats: Minors' Perspectives on Disclosing, Reporting, and Blocking.

"Thank you for talking to me earlier about what I can do about this revenge porn situation. **Staying anonymous really is the most important thing to me,** I don't need more people, and definitely not my parents, finding out about this. **The Victim Support website was really reassuring about confidentiality, and I am going to use Report Remove to get my pictures taken down**." Call to Childline from a girl, aged 16.

This measure would therefore reduce a key barrier to reporting for children and young people by ensuring there is clear information provided at the outset about next steps after a report is made. This can help to reduce fears about their anonymity, provide reassurance that if they have experienced an illegal harm it is not their fault and they are right to report it, and explain what action the site may take against the reported user, as well as signposting to appropriate support[69]. Improving transparency is a reasonable expectation for services, who should already be providing information about how reporting works.

5B. Having an easy to find, easy to access and easy to use complaints system

We support this measure. The suggestion that users should be able to provide contextual information (A5.4 (d)) is particularly important. Young people consistently raise that despite material being illegal, it is not taken down when reported. Girls that the NSPCC work with have raised this issue, noting that when they have reported semi-nude images of their peers online, they have regularly not been removed. It is crucial that there is space to explain why an image must be removed to speed up the process of taking it down and ensure that children do not need to make repeated reports and complaints.

5C. Appropriate action – sending indicative timelines

We strongly disagree with the decision not to recommend that services update users on the outcome of their reports and complaints. There are a number of reasons why this is crucial for children.

Firstly, if a child has made a complaint about an illegal harm which is impacting them, such as grooming, it is vital that they know if action has been taken against the perpetrator. This will significantly impact their sense of security and their freedom to continue to use the platform.

Secondly, not knowing the outcome means children are unsure if they need to take further action or if the issue has been resolved. Childline hears from children who have made a report but, because they do not know the outcome, want to know if there is more action that they need to take. By providing information about the outcome, children will have a much better understanding what has happened and can be signposted to further support if that is appropriate.

"I don't know what I can do besides stalling for time. This account is demanding money from me after I sent them nudes. They won't stop messaging and calling me. I haven't sent any money, I don't have any to send. **I've reported the account but what else am I meant to do?**" Call to Childline from a boy, aged 15.

Thirdly, not knowing if anything has happened with a report reinforces the opaque nature of reporting and puts children off using these systems again because they do not know if they have made a difference. Young people working with the NSPCC raised that it is reassuring to know what the outcome of the report is and means they are more likely to use that reporting system again.

[69] All signposting to external support should be done in consultation with the service to ensure it is equipped to manage any new demand.

Research reinforces this, highlighting that one of the main reasons that reporting is seen as ineffective is due to the lack of information provided about the outcome.[70]

We strongly recommend, therefore, that this measure is changed to include a requirement that users are updated on the outcome of their report. As a minimum, users that make a report about CSEA must receive an update, given the severity of this crime. Alternatively, the measures regarding default settings to children could be amended to include a requirement that children are always updated on the outcome of a report, to help develop good reporting behaviours. But it is critical that this change is made, as otherwise there is a real risk that the other measures aimed at strengthening reporting will be undermined because children will ultimately not know if reporting has changed anything.

5I. Dedicated reporting channels

We strongly recommend that in future Codes a new measure is included that providers with a medium or high risk of CSEA establish and maintain a dedicated reporting channel for trusted flaggers.

The NSPCC operate a Trusted Flagger process where we have a direct link with some online service providers. This is generally used when members of the public share content directly with either of our counselling services, Helpline and Childline, to request take down of content or voice concern about the content. It enables us to feed into a service's moderation system and ensure they are prioritising the most dangerous and severe material for removal.

There have been some challenges with this process as online service providers often require an undue level of information to action reports, or content has not been removed despite a trusted body having raised it.

Requiring the creation of a dedicated reporting channel in the Codes of Practice and setting out key principles this should meet to ensure its efficacy would therefore be highly valuable for strengthening these systems by reducing the burden on trusted flaggers and ensuring their reports are meaningfully actioned. This is crucial for ensuring that moderation is informed by external experts and that high priority material can be identified and removed. We recommend considering the Guidance for Trusted Flagger Programmes produced by UKCIS to inform future measures.[71]

Terms of service and Publicly Available Statements

Question 29: Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views.

Question 30: Do you have any evidence, in particular on the use of prompts, to guide further work in this area?

Default settings and user support for child users (U2U)

To inform our response to this section, we worked with a group of for young people who provided feedback on the measures proposed and shared their ideas about disrupting grooming online. Their input has shaped our answers, and we also refer to them directly as 'our young people's panel'.

---

[70] Vilk, V. and Lo, K. (2023) Shouting into the Void: Why Reporting Abuse to Social Media Platforms Is So Hard and How to Fix It; Luria, M. and Scott, C. F. (2023) More Tools, More Control: Lessons from Young Users on Handling Unwanted Messages Online.

[71] UK Council for Internet Safety. Trusted Flagger Programmes: Guidelines and Best Practice.

Question 31: Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views.

We support the measures proposed in this section. In this answer we set out why we support the approach, some challenges and limitations, and then provide specific feedback on the individual measures proposed.

It is vital that friction is introduced into the grooming pathway, and children are empowered to have better control over their privacy and safety online. At present, it is far too easy for children to be contacted by strangers online which then leads to grooming and child sexual abuse. There is overwhelming evidence from Childline and a wide range of sources that direct messaging in particular is used by perpetrators to make contact with children. It is therefore very welcome that it will be more challenging to identify and contact children, especially via messaging services.

There are, however, limitations to this approach. There is no silver bullet to tackling grooming online. These settings will need to work alongside other proactive detection measures and tools which target perpetrators. The burden cannot be on children to have to protect themselves from abuse. Making these options default but with the option to turn them off rightly empowers children, but also means they will ultimately need to make decisions about how they behave online and where they want to trade off their privacy for less restrictive experiences. We urge that Ofcom uses its information gathering and supervision powers to better understand the range of ways that platforms can disrupt grooming so that a wider range of technical solutions can be introduced that take the burden off children and ensure safety is embedded for all.

Another limitation is that effective age-checking will be fundamental for these tools to work. Children will only have these settings on by default if they are recognised by the service as under 18 years old. When discussing these measures, our young people's panel immediately questioned how well they would work in practice, given so many young people do not sign up with their real age when making social media accounts. Ofcom's own research has found that almost half of children aged 8-15 have at least one social media profile with a user age of at least 16, and that a third of those aged 8-17 had a profile with a user age of at least 18.[72] With that knowledge, it is therefore vital that Ofcom ensures future Codes set out rigorous measures for how platforms improve their age assurance processes to ensure children are accurately identified and protected. This will also help tackle the creation of fake profiles by perpetrators, as discussed in detail in Q16.

Connected to this, if platforms are going to implement these measures once the final Codes are introduced, they must also be required to consider any steps they can take to identify existing accounts which have put the incorrect age upon signing up. Without retrospectively identifying children already on a service, not all young people will be able to benefit from these settings and they will face a disproportionate risk.

7A. Safety defaults for child users.

A7.2 If the service has network expansion prompts or connection lists, the provider should implement default settings ensuring that…

a) child users are not included in network expansion prompts presented to other users;

b) child users are not presented with network expansion prompts;

[72] Yonder Consulting (2022) Children's Online User Ages Quantitative Research Study. Ofcom.

c) child users are not included in the connection lists of other users;

d) connection lists of child users are not displayed to other users.

We support these measures as ways to make it more challenging for perpetrators to quickly identify large groups of children. It is clear that network expanding prompts are particularly risky for children. Children are often encouraged to quickly expand their connections which can lead to unwanted interactions with adults – an example of how a service's business model, relying on high user engagement, can put children at risk.

"I'm feeling a bit weirded out right now. You know on Snapchat how you can just add people on Quick Add? So, I added some people my friends' knew cos it said they had mutual friends. Then one of them replied to something on my story saying I was 'hot' and I had a nice figure. At first, I was like thanks, then I asked how old he was and the man said 22?! I'm like WHAT - and blocked him like that. Don't you think that's weird telling a 13-year-old they're hot?!" Call to Childline from a girl, aged 13.

However, for points (c) and (d), our young people's panel raised that there are a number of benefits to being able to see the lists of child users they are connected with. They noted that this is helpful for checking if someone who adds you is friends with your friends. Knowing that you are connecting with someone who is a mutual contact, rather than someone with no connection, helps them consider if it is someone they should add.

As noted above, there are trade-offs that children will have to make when deciding if they want to keep these settings on. They will need to weigh up the safety benefits against the other benefits of these functions. Whilst we support these measures remaining in the Codes of Practice, it is an example of the limits these settings may have in the eyes of children and the importance of taking a holistic approach to tackling grooming.

A7.3 If the service has direct messaging, the provider should implement default settings…

We strongly support these recommendations for empowering young people and giving them more control over who they interact with online. These settings were identified as the most important ones by our young people's panel, who noted that the ease with which strangers can send inappropriate messages to children is one of the biggest issues which needs to be tackled. In particular, they highlighted the importance of being able to more easily reject interactions that they do not want. They wanted this to be built into the messaging functionality, and ensure they do not have to see the messages before doing this. These settings would effectively address these needs.

Our panel also raised the importance of being able to see who a user is, before deciding whether to accept a message from them. Services must provide simple ways for a child to check who has messaged them before they proceed and accept a message – this could be included in the Code as an extra recommendation for direct messaging.

7B. Support for child users

More safety support information for children will improve the efficacy of new settings and help children to make informed decisions. Ultimately, the success of these support messages will be determined by how they are designed, which will likely differ for each platform. Services must engage directly with children or representative groups in order to develop messaging that is useful and accessible. Our young people's panel emphasised that whilst these sorts of messages would be helpful, at the moment this information tends to be buried in long policies or updates that can be

easily skipped. A new approach must be taken by services if this support is to reach and help children.

We recognise the benefits of not including specific requirements for this support information in the Codes. However, it will be complex to develop. Broadly, this messaging must avoid jargon and ensure the tone is not overly formal, without being patronising or 'cool'. It will need to balance empowering people to take steps to protect themselves, without putting the onus on children and young people to solve situations themselves or imply it is their responsibility to do so. Where possible, it is important to give as much information as possible about what will happen next, and different ways of accessing further support.

The recommendation that services should provide information to child users about the risks of disabling a default setting, and assist them to understand the implications, will be a particularly important one to get right. After a child has tried to disable a setting and received this message, they should be re-prompted to check if they still want to turn the setting off. It will also be important for services to consider that alarmist messaging and a focus on the dangers of being online is likely to be off-putting to young people.[73] For many children and young people, their online lives are an integral and positive part of their life, and overly negative information is unlikely to align with their understanding of the online world. Support information about settings need to recognise the benefits of being online and the importance of children's safety and privacy when explaining the advantages of these settings.

It would be valuable for companies to have guidance to help them in developing effective and age-appropriate information. Ofcom could produce this guidance themselves or commission it.

Our young people's panel also proposed a number of ways that support information could be developed and shared so that it is useful for children. This included:

- Provide reassurance that the child can always changes the settings back, and share some information about how they can do this.
- When there are updates or reminders about new settings, this information could be shared when a user opens the app, in messages from the platform, or in promotions from the platform.
- Key messages should have a minimum amount of time that the child has to look at the information for before they can scroll or skip it.

One way to ensure that these services are child-centric would be to undertake 'user-testing' with a panel of children and young people. Services would need to ensure comprehensive safeguarding and support systems are in place to undertake this. However, especially for large platforms with a high number of child users, this would add significant value in ensuring the support information covers the right issues, is appropriately framed, and shared in the best places for children.

When reviewing a service's compliance with this measure, Ofcom should not just check whether this information is technically available, but whether it is accessible and child-centred.

Question 32: Are there functionalities outside of the ones listed in our proposals, that should explicitly inform users around changing default settings?

[73] Beckett, H., Warrington, C. and Montgomery Devlin, J. (2019) Learning about online sexual harm. University of Bedfordshire: Commissioned and undertaken on behalf of the Independent Inquiry into Child Sexual Abuse.

We welcome the measures around direct messaging. They could be strengthened by implementing settings that help detect and warn children about potential illegal activity in messaging services. ***In particular, we recommend that in future Codes of Practice, platforms are prompted to develop tools that enable client-side nudity detection and mean potentially explicit images are blurred***, with the user deciding whether they want to view, delete, or report the image before viewing it. This setting is used by Bumble[74] and Apple[75], introducing important friction to grooming and image-based sexual abuse pathways. This is also an example of why it is important to move from grounding the Codes in widespread industry practice to promoting more innovative approaches. These image-blurring tools are not yet widely used, and so it would be an effective way for Ofcom to raise the bar for best practice.

"Some random guy sent me a picture of his penis on WhatsApp. I get these things all the time but that doesn't make it okay, or that we should just put up with it. I'm extra annoyed it was over WhatsApp cos the pictures are now automatically saved on my phone." Call to Childline from a boy, aged 14.

There are other default settings that would reduce the risk of children's accounts being identified and targeted by offenders. Children have highlighted that they would like to be less discoverable online, with the option to make their profile more public if they want to.[76] ***Two key changes that would support this and should be introduced in future Codes are:***

- ***Tagging***: For services which enable tagging, services should turn off visible tags. This would mean that other users cannot view the accounts of users tagged in content shared by children.
- ***Information in bios***: Platforms which have free text bios often contain personal information about a user, potentially including where they live, their age, and their interests. This information should not be visible to users that a child is not connected with.

One member of our young people's panel raised that they often get notifications about the activity of other users they are not connected with, including adults. Having even just one mutual connection can be enough to lead to regular notifications about users you do not know on some platforms. They raised that this can encourage young people to engage or connect with strangers online. They recommended that as well as not sharing network expansion prompts with children, there should be a default setting which means children do not receive notifications about accounts that they are not connected with.

Question 33: Are there other points within the user journey where under 18s should be informed of the risk of illegal content?

As mentioned above, effective age checking processes are fundamental to ensuring these settings protect children. Our young people's panel all emphasised that it is currently incredibly simple to create an account with the wrong age. They suggested that ***prompts or support information at the stage of creating an account*** might help children consider the risks of putting in the wrong age and the protections being recognised as a child affords. This information could also support children to understand why they are being asked to provide personal information that feels private to them.

[74] Bumble. Unsolicited Lewd Photos: Why am I seeing a blurred image?
[75] Iovine, A. (2023) Apple's Sensitive Content Warning will combat cyberflashing. Mashable.
[76] Luria, M. and Scott, C. F. (2023) More Tools, More Control: Lessons from Young Users on Handling Unwanted Messages Online.

We also recommend that, for children's accounts with at least one default setting turned off, services set ad-hoc prompts encouraging the child to review their settings and providing information about how to turn safety settings back on.

Typically, users only get prompts when first setting up an account and if they choose to make changes. This puts the onus on child users to reconsider their safety settings in the future, when instead the platform could make timely suggestions throughout the life cycle of the account. These prompts could be periodically (e.g. annually), once the user has a certain reach, or when they start to engage with a new feature on the platform. These messages would need to be tailored and balanced in frequency to prevent message fatigue.

Recommender system testing (U2U)

Question 34: Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views.

This proposal should be extended to apply to all large, multi-risk services, regardless of whether they already have on-platform testing of recommender systems in place. Whilst this will pose a cost to those who do not have one in place, we believe this is proportionate given that this is a key safety by design measure.

Without undertaking robust safety testing, services will not be able to fully understand the impact of changes to their recommender systems. Given the reach and impact of large, risky services, it is vital that they understand the consequences of any changes to their algorithms before making widespread change. This will ensure they are actually designing their systems to be safer from the outset, based on testing and user insight, rather than retrospectively trying to adapt system changes after they have already been rolled out and potentially caused significant harm.

Research and evidence clearly demonstrate that there is a connection between algorithms designed without user safety in mind and harm to children. The Centre for Countering Digital Hate found that a 13-year-old child could be recommended suicide content within 2.6 minutes of being on TikTok.[77] Recent research from the Molly Rose Foundation found that algorithms enable dangerous content to have a far reach; over half of the most engaged harmful posts (relating to suicide and self-harm) they surveyed on TikTok had been viewed over 1 million times.[78] On X, recommender systems have shared videos of children being sexually assaulted and prompted users to follow content relating to exploited children.[79] Young people have reported to Childline that they have been recommended child sexual abuse material online:

"Twitter has been recommending me posts about a manga cartoon series I'm really into, however some of the posts really unnerve me. Some show the students from the show in sexual scenarios with their teachers or with other students. The students in the show are, like, 15-16, which makes me uncomfortable because they're minors." Call to Childline from a girl, aged 18.

[77] Centre for Countering Digital Hate (2022) Deadly by Design: TikTok pushes harmful content promoting eating disorders and self-harm into users' feeds.
[78] Molly Rose Foundation and The Bright Initiative (2023) Preventable yet pervasive: The prevalence and characteristics of harmful content, including suicide and self-harm material, on Instagram, TikTok and Pinterest.
[79] Keller, M. and Conger, K. (2023) Musk Pledged to Cleanse Twitter of Child Abuse Content. Is It Working? New York Times.

Research testing the type of content recommended by algorithms using simulated accounts for child users has found that children's accounts with indicators of vulnerability are particularly likely to be recommended potentially illegal suicide and self-harm content, as well as eating disorder content.[80]

Analysis also suggests that Facebook's algorithms have recommended users who are in private offender groups to similar sites, through determining the common characteristic of interest (a sexual interest in children).[81] This even includes recommending similar groups in multiple other languages.[82]

There is a strong consensus about the risks posed by algorithms designed without user safety in mind. The principles of safety by design demand that companies are building protections for their users into the fabric of their services – rather than just retrofitting safety measures to fundamentally risky services. If large services do not already have systems in place to test the impact of changes to their recommender systems, then that is an indication of weak safety governance. It is not an acceptable reason for continuing not to test, and the Codes of Practice should rectify this.

Question 35: What evaluation methods might be suitable for smaller services that do not have the capacity to perform on-platform testing?

Question 36: 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. Are you aware of any other design parameters and choices that are proven to improve user safety?

Enhanced user control (U2U)

Question 37: Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views.

We agree with the proposals in this section, particularly around user blocking, muting, and disabling comments. However, we are concerned some of the analysis in this section significantly overstates the benefits to users, and particularly to children.

For example, it is noted that grooming perpetrators create fake user profiles to appear less threatening to children. It is then suggested that blocking / muting options will help protect users from illegal content because 'a user who could predict that they would be targeted with illegal content… would be able to prevent it from happening'. And it is argued specifically that 'providing child users with the option to block or mute users would be an effective tool to enable them to protect themselves online, particularly in relation to grooming and other child sexual abuse offences'.

This analysis completely misrepresents the nature of online grooming and child sexual abuse. The reason that perpetrators create fake profiles to groom children is because they are deceptive and misleading and mean that children are more likely to trust them. Evidence from Childline shows that when children accept connections with (what turn out to be) fake profiles, they initially think they are someone their age – they do not immediately identify a risk which leads them to block. Once the offender has begun to groom a child, they often quickly use image-sharing, exploitation, and blackmail. Where this then makes a child uncomfortable, many feel unable to report the abuse or

[80] Bryce, J. et al (2023) Evidence review on online risks to children. London: NSPCC.
[81] Putnam, L. (2022) Facebook Has a Child Predation Problem. Wired.
[82] NSPCC (2022) Time to Act: An assessment of the Online Safety Bill against the NSPCC's six tests for protecting children.

block the perpetrator because they are being threatened – including threats to have their nude images shared, and threats of violence.

"I really thought this guy I was talking to was my age. We had been texting for a while, then it became sexting and then he asked for nudes. I sent some but he insisted I send more with my face in. He just kept asking and peer pressuring me so I just did it. Now he's told me he's 32! I don't want to talk to him anymore, but he has my pictures and is threatening to share them." Call to Childline from a girl, aged 16.

"I thought the person I was talking to was a lady and she asked to see my privates. She said she was 16, now she's saying she's 9 and that I've broken the law! She says I have to send her my mum's card details or she'll have me arrested. I'm so scared of getting in trouble with my mum and the police" Call to Childline from a boy, aged 13.

For the avoidance of doubt, we are not opposed to these measures, but we strongly disagree with the suggestion that they provide a meaningful safety net for children against grooming. The implementation of these measures should not enable services to evade implementing safer design features, proactive detection tools, and strong moderation systems.

Question 38: Do you think the first two proposed measures should include requirements for how these controls are made known to users?

These measures should be promoted to users. Given children's apathy to reporting systems generally, they are unlikely to be proactive in seeking out other similar tools without being prompted.

Our young people's panel raised that blocking tools are often hidden and not straightforward to use. One girl raised that she would not know how to block someone on WhatsApp, for example, or if it is even possible. Young people raised that it would therefore be particularly useful to be notified about blocking / muting options in direct messages, as this is where they are most likely to receive unwanted contacts. ***We recommend that services are required to make these controls known to users, and that this should be informed by where these measures will be most useful to users.***

Question 39: Do you think there are situations where the labelling of accounts through voluntary verification schemes has particular value or risks?

User access to services (U2U)

Question 40: Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views.

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

Question 41: 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)? What are the advantages and disadvantages of the different options, including any potential impact on other users?

We recognise the challenges of blocking users who have shared CSAM. Offenders are sometimes able to continue to make new accounts to share content. Even if a specific IP address is blocked, it is likely an offender would be using a VPN and so evade detection.

However, it is not acceptable that adult offenders who have shared CSAM on a platform would face no repercussions. Moreover, there may be a deterrent value for low level offenders if there is a clear risk of being blocked and reported to the police.

This is an area where information from tech companies on the efficacy of blocking tools and their current practices would help the development of appropriate, well-targeted recommendations. Systems which block IP addresses and report users to law enforcement should be considered in particular, to prevent the offender from causing further harm on other platforms. ***Ofcom should use their information-gathering powers to understand how regulated services are approaching blocking users who have shared CSAM to inform future iterations of the Code.***

The strength of service's blocking tools will in part depend on other measures that a service has in place which are currently missing from this Code of Practice. For example, measures which limit how many (fake) profiles a user is able to make would mean it is easier to target blocking. Cross-platform cooperation would enable services to share signals about accounts which have shared CSAM. ***We discuss these measures further in answer to Q16, and recommend that Ofcom considers the connections between these issues in the next Code of Practice.***

Question 42: How long should a user be blocked for sharing known CSAM, and should the period vary depending on the nature of the offence committed?

Criteria for blocking a user for sharing CSAM should be developed to help services implement this proportionately and consistently. This guidance should consider the age of the offender, the nature of the CSAM, the intention behind sharing and whether it was a repeat offence. We offer some considerations on two of these areas below, however we recognise that this is a challenging area and would welcome further discission with Ofcom ahead of future Codes.

<u>Age</u>

As in the offline world, children should be treated with greater leniency in cases where they have shared imagery. Children may be more likely to share CSAM out of outrage, shock, or misplaced humour, without malicious intent and without considering the severity of this. It is also vital that children do not face any penalties for sharing their own images online.

<u>Intention</u>

There are cases where adult users share CSAM out of outrage, shock, or to try and get a post taken down. This material must always be instantly removed and the risk to children means that in the first instance, a service should implement its standard blocking procedure. Services may decide to have a system, however, to provide users with the option to appeal their ban. Where there is reasonable evidence that they did not mean to commit a crime – e.g. it is Category C imagery, they posted a comment which called on law enforcement to remove the image, and they have only done this once – services may consider repealing the ban.

Question 43: 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. 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?

More societal damage is done by not safeguarding a child in an abusive situation than the damage caused by a false positive from an automated decision which is investigated by human moderation and subsequently dropped. Automated systems are overwhelmingly more likely to detect genuine

CSAM than false positives, especially for known CSAM. As noted above, NCMEC estimate that the hash matching tool PhotoDNA has a false positive rate of under one in one trillion.[83] Moreover, automated tools are typically used alongside human moderation which limits the likelihood of false positives and ensures benign images are not misreported.

Blocking measures must therefore not be disproportionately weakened. To manage this risk, Ofcom could recommend that services ensure human moderators are involved in cases where a user might be banned from a service. In these cases, users should either be temporarily banned or have key features disabled (e.g. posting and direst messaging functionalities) whilst their offence is being reviewed by a human moderator, to limit risk in the interim.

Ultimately, implementing accurate automated detection systems, supported by efficient reporting channels, will reduce the risk of law-abiding users being negatively impacted. This approach will limit the likelihood of false positives and enable users who feel lawful content has been erroneously classified to access speedy redress options. Efficient redress is a proportionate way of mitigating this risk whilst still fulfilling the overriding imperative to stop the spread of CSAM.

Service design and user support (Search)

Question 44: Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views.

We strongly support the proposals to provide users with warnings / support information if they are searching for CSAM or suicide content (7A and 7B).

The requirement that services develop and maintain an appropriate list of terms commonly used to search for CSAM provides a potential opportunity for cross-platform collaboration. Over time, we would expect services to identify new terms which perpetrators are using to search for and access (particular forms of) CSAM. ***In future Codes, services should be required to share these new terms with other services, or with a central provider, to ensure that efforts to tackle access to CSAM online are consistent and comprehensive.***

Signposting users to appropriate organisations and resources when they are searching for illegal / dangerous material are important measures which we support, but it is important that this process is done with the consent of and in collaboration with the organisation being signposted to. Ofcom should also consider how income generated from fines in the future could be used to help fund these support organisations, particularly those providing helplines and services which require significant resources.

The provision of suicide crisis prevention information must be appropriate for both child and adult users, with helplines or organisations which will be recognisable for children and young people included in signposting.

Cumulative assessment

Question 45: Do you agree that the overall burden of our measures on low risk small and micro businesses is proportionate?

---

[83] Steinebach, M. (2023) An Analysis of PhotoDNA. In *The 18th International Conference on Availability, Reliability and Security (ARES 2023), August 29--September 01, 2023, Benevento, Italy.* ACM, New York, NY, USA 8 Pages. https://doi.org/10.1145/3600160.3605048.

Question 46: 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?

We agree that these measures are certainly not overburdensome. As highlighted in our response, there are a number of areas where further measures should be applied to these services to ensure they have the governance and functionalities to meaningfully protect children.

Since 2017, analysis of police FOI data indicates that over 150 different apps, games and websites were used to target children for grooming offences.[84] It is highly unlikely that all of these will be classed as large platforms. It is therefore vital that smaller platforms with a risk of child sexual abuse are held accountable for improving their services. Whilst the Act recognises the importance of proportionality, its core aims (set out in Section 1) do not distinguish between services. All regulated services are expected to safe by design, and to afford a higher standard of protection to children. This will only be possible if reasonable safety measures are imposed on small services that pose a specific risk to children's safety online, and is not something we think is fully delivered in this first Code.

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

As with the measures for small services, we agree that the plans for large services are not overburdensome. The resource, expertise and reach of these services mean the current measures are proportionate and should be delivered at pace.

We note that many of the measures proposed in the Codes are already widely used by large services. This includes the recommendations for hash matching, reporting and blocking functionalities, and published terms of service. This indicates that it is reasonable to expect these to be rolled out as standard in the industry. It also suggests that, alone, these measures will be insufficient in tackling CSEA online. In the future, services must be required to implement new features and tools that mark a step change in how they protect children. We have recommended such measures throughout our response, and they include using more proactive tools to detect and disrupt abuse, making changes that will target perpetrator behaviour, and being much more ambitious with requirements for private messaging services.

Statutory tests

Question 48: 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? If not, why not?

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

The Illegal Content Judgements Guidance (ICJG)

Question 49: Do you agree with our proposals, including the detail of the drafting? What are the underlying arguments and evidence that inform your view?

This guidance is complex, given it covers long-standing offences, new ones, and others where the law is unclear. Given the legal nature of this document we are not best placed to comment on the

[84] Based on NSPCC Freedom of Information requests to UK police forces from 2017/18 to 2022/23.

approach and proposal. However, we would refer Ofcom to the Online Safety Act Network's response which sets out some challenges with this section.

Investigating suspicious activity

We suggest greater consideration is given to circumstances where a reasonable suspicion of illegal content / activity should be sufficient for a service to take action. In cases where a service has received suspicious signals indicating illegal activity, for example, a new adult male account has tried to connect with hundreds of girl's accounts at random, this should trigger an investigation to determine if illegal activity has taken place. This is particularly important for any suspicions of CSEA – content should not have to clearly meet the legal threshold to justify further investigation or action in these cases.

It should therefore be clarified that services can, and should, act on suspicious signals that indicate a child is at risk, using evidence from automated services and further investigation by human moderators. This approach could be supported by providing further detail in the content moderation standards about at what threshold of suspicion a service should take action to determine illegality.

Regardless of whether an online service removes content because it is determined to be illegal, or because it infringes their terms of service, Ofcom should ensure that online services report any suspicious activity through the appropriate safeguarding channels. There is a risk that services which use their terms of service to take down content or ban users, rather than considering in the round whether activity is illegal, may not follow through with a full safeguarding response and make the necessary reports to law enforcement. This is especially relevant for facilitation where activity may be more nuanced but still have safeguarding implications.

Facilitation

Another area that needs clarification is how services should approach content which facilitates abuse.

In the section on CSAM, it is suggested that content should be considered illegal if it does not itself contain illegal imagery, but links or otherwise directs users to how this material can be found or created. We welcome the inclusion of some detail of this, but considerably more information and examples should be provided to help services identify facilitation content. The usage examples focus on hyperlinks, which are a clear way that facilitation can be carried out. However, there are more complex examples which should be classed as facilitation but may not be removed by services.

For example, there is growing evidence about the harm caused by so-called 'tribute sites', in which offenders create online profiles that misappropriate the identities of known survivors. These fraudulent accounts, which typically adopt survivors' names and feature non-harmful imagery at the account or profile level, are then used by offender communities, including to signpost to CSAM on the dark web.[85] Another example is that offenders have used Instagram accounts to aggregate legal images and videos of young girls, leading to offenders meeting in the comments to signpost each other to offender communities on other apps where they can share CSAM.

Given the intention of these and similar pages is to assist the commissioning and committing of child sexual abuse, they must be addressed as a form of facilitation and removed by services. This is

[85] WeProtect data generated by Crisp Consulting

explicitly called for in the Act, which says services must mitigate and manage the risk of a service being used to commission or facilitate a priority offence.

Either in this document or in separate guidance, we strongly recommend that Ofcom shares further detail about the ways that CSEA can be facilitated online, providing more nuanced examples about the content services are expected to remove. Ofcom should also publicly outline when they will include measures to tackle facilitation in Codes of Practice.

Question 50: Do you consider the guidance to be sufficiently accessible, particularly for services with limited access to legal expertise?

Question 51: What do you think of our assessment of what information is reasonably available and relevant to illegal content judgements?

CSAM

One of the most important types of reasonably available information noted in this guidance is information provided by a subject in a reported image. If the complainant is saying that they are featured in the imagery and they are a child, then it should be removed and reported as CSAM. We know that this often does not happen when children report nude imagery, particularly when it is older teenagers. One young person who worked with the NSPCC raised that despite explicitly stating in their reports to Snapchat that nude images they reported were of someone under-18 (their peer), these images were not removed. This issue is also evident in calls to Childline. It is vital that services use and prioritise the information being provided in a report to determine if it is CSAM and report it accordingly.

"I found out my child had been groomed by an older child. There were multiple Instagram accounts of her, she'd been forced to pose in suggestive ways. I reported it to Instagram, and I got absolutely no response. Some would go over time, but other accounts were still up. It wasn't until I contacted the police that all the images got properly taken down but some of the accounts are still there. I want them gone." Call to NSPCC Helpline from a parent.

"I am feeling sick with fear. I was talking with this guy online and trusted him. I sent him quite a lot of nude pictures of myself and now he is threatening to send them to my friends and family unless I send him more nudes or pay him. I reported it to Instagram but they haven't got back to me. I don't want to tell the police because my parents would then know what I did and would be so disappointed" Call to Childline from a girl, aged 14.

It would also be valuable for the ICJG to acknowledge the connection between CSAM offences and the new offence of sharing an intimate image without consent. There will be cases when services find determining the age of a user more challenging. For example, if an account indicates the complainant is over 18, but the user is reporting the images as CSAM. However, the introduction of the new intimate image abuse offences means that providing the image has been shared without consent, it must be removed. Companies must still make every effort to determine if the image is CSAM, so it can be reported to the appropriate channels, but the combination of these offences means the legal basis for removing these images as a matter of urgency is clear.

"I need your help – it's really urgent! There's currently an Instagram account of me and it has my nudes and I want it to be taken down as soon as possible! I've already reported it to Instagram more than 10 times now but it's not deleting. I don't know what else I can do. Please can you get it taken down because it's affecting my mental health!" Call to Childline from a female, aged 18.

We welcome that Ofcom has recognised the threat of AI-generated CSAM in this guidance, and notes that illegal content could include discussion of how to use generative AI to make illegal content and sharing links that directly facilitate the production of artificial CSAM. Ofcom should continue to monitor how AI-generated CSAM is being facilitated or commissioned through regulated online services, and work with services to ensure this is being removed as illegal content.

Grooming

We strongly agree with Ofcom's view that 'a potential victim of grooming, who declares themselves to be a child, should usually be believed'. It would be incredibly challenging to make a false report, due to the content that would have to be manufactured. Children already face considerable barriers in reporting and getting support from services when they have experienced illegal harms. Ensuring they are trusted by platforms and the burden of proof is not disproportionately high is therefore critical.

One area which demands further consideration is how child sexual abuse in virtual reality (VR) environments should be addressed by services. Child sexual abuse in VR can blur the boundaries between different offences. For example, NSPCC-commissioned research found that VR multi-user spaces provide opportunities for offenders to commit child sexual abuse against a child in VR and engage the child in child sexual exploitation.[86] Accurately identifying and reporting illegal activity in VR environments will therefore be complex and require investment in expert moderation teams.

The nature of reasonably available and relevant information in these environments will also look significantly different to standard social media platforms, particularly due to the different ways that content shared by users is captured and recorded. Services currently record audio and video interactions in VR environments for different lengths of time to enable users to make reports – Meta's Horizon Worlds[87] records the last few minutes of a user's most recent audio, video and other interactions and Fortnite's[88] voice reporting system means that the last five minutes of voice chat is recorded. Ofcom will need assess how long data should be recorded and stored to allow illegal harm to be reported by users and addressed by moderators, ensuring the persistent nature of the metaverse does not mean it is easier for perpetrators to evade detection.

Ofcom should consider if further guidance is required which sets out how illegal activity should be identified and judged in immersive environments, and what information services must store, in a safe and privacy-preserving manner, to assist with the identification and reporting of illegal activity.

It may also be helpful to include more usage examples in the ICJG which are directly applicable to immersive environments.

Illegal suicide and self-harm content

We are in agreement with the assessment in this section. One area which may require further guidance for services is how to approach content relating to suicide which is posted by people who are themselves in significant distress. There may be circumstances, such as posts in forums related to suicide and self-harm, where a user's content is a risk to other users but also indicates that they are suffering. In these circumstances, users must be signposted to further support. For example, if a child has a post relating to suicide removed and receives a notification from the service that they

[86] Allen, C. and McIntosh, V. (2023) Child safeguarding and immersive technologies: an outline of the risks. London: NSPCC.
[87] Meta (2023) Supplemental Meta Platforms Technologies Privacy Policy.
[88] Epic Games. How does voice reporting in Fortnite work?

have broken community guidelines, without any recognition of the support the user might need, this risks further isolating the young person at a time when they may be highly dependent on online spaces for support and engagement. ***Services and Ofcom should therefore consider how vulnerable users can be signposted to external support and resources when content relating to self-harm and suicide is removed.***

Fictional descriptions and the glorification of suicide and self-harm are both noted as out of scope of these offences because it is unlikely that intent can be inferred. Whilst we recognise the limitations of classifying this content as illegal, we want to reinforce the major risk that this content can pose to children and young people, particularly when viewed at a high volume. We expect this issue to be returned to in drafting of the Children's Safety Codes.

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

Information powers

Question 52: Do you have any comments on our proposed approach to information gathering powers under the Online Safety Act?

As is recognised in Volume 6, information gathering is critical to the implementation and success of the new regime. The lack of transparency from the tech sector has created a severe information imbalance, leaving civil society and researchers unable to fully understand the mechanisms which drive harm for children online. We urge Ofcom to promote a culture of transparency through the information gathering powers – this must stretch beyond only companies and Ofcom holding information about the risks to children online and the solutions to tackling these harms. Ofcom must ensure that civil society and researchers can access information to evaluate the steps taken by services, highlight where change is needed, and make full use of the supercomplaints mechanism.

The information gathering powers should be used as early as possible to help strengthen the Codes of Practice. This is particularly important for addressing some of the key gaps in this first Code – such as measures to identify unknown CSAM and target perpetrator behaviour. As well as using information notices to establish different parts of the regime, Ofcom should also commit to understanding what innovative methods services are using to address these issues and gather evidence on the efficacy of these tools ahead of the next Code.

We are concerned with the suggestion that in some cases, a service (rather than Ofcom) could be required to appoint a skilled person to make a report before Ofcom issues a technology notice. It is not clear how Ofcom will ensure that the service is appointing a skilled person who will not be bias in favour of the service and their priorities. We strongly suggest that services should not be able to select the skilled person. Platforms are likely going to resist Ofcom's use of the proactive technology powers, and it is vital that an external experts involved in this process are impartial and not tied to or supported by the service in question. As a minimum, Ofcom should provide further information about the criteria a skilled person will need to meet and how they will ensure any skilled person will not be bias towards the platform's interests.

Enforcement powers

Question 53: Do you have any comments on our draft Online Safety Enforcement Guidance

There is limited detail on the scenarios in which Ofcom would expect to use the enforcement powers, which we would like to see further defined. Services must understand in what

circumstances enforcement powers will be used if they are to have their intended purpose of incentivising good practice and deterring non-compliance from the outset. There may be cases of egregious non-compliance early on (such as services explicitly stating they will not be complying with the regulation), where the full weight of Ofcom's enforcement powers will be required.

One key area that must be clarified in this guidance is how Ofcom will hold senior managers liable for compliance with confirmation decisions relating to CSEA requirements. Whilst the guidance acknowledges that it is a criminal offence to fail to comply with a CSEA requirement imposed in a confirmation decision, no detail is provided as to how this power might be used in practice.

Such little detail about the scope of senior manager liability in this guidance risks indicating that it is not a power Ofcom will look to use. A core purpose of this particular power is to ensure that senior managers give full consideration to their compliance duties and prioritise children's safety in decision making. If it appears that this power will in reality never be used, they are unlikely to feel as strong an impetus to ensuring regulatory compliance.

In particular, it should be clarified that services must name a senior manager / senior mangers in response to a confirmation decision. These managers will be responsible for the service's response to any requirements in the decision and can be held liable if the service fails to meet the requirements.

Annex 13: Impact assessments

Question 54: 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?

Question 55: 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.