NSPCC Response: Ofcom's Consultation on Protecting Children from Harms Online

July 2024

Volume 2: Identifying the services children are using

Children's Access Assessments (Section 4).

Do you agree with our proposals in relation to children's access assessments, in particular the aspects below. Please provide evidence to support your view.

1. Our proposal that service providers should only conclude that children are not normally able to access a service where they are using highly effective age assurance?

We agree with this approach.

Children access a wide range of services online, including those which are not intended for their use. As well as this, services have largely failed to restrict access to under-age users due to weak age-assurance systems. [1] This means that children have been able to seek out and easily access services which are not built with their safety in mind. This causes significant harm to children, including children being aware that what they are seeing is harmful to them, but being unable to stop engaging with the content. Without the implementation of highly effective age assurance measures which stop under-18s from accessing adult sites, this will continue.

"Since the age of about 8, I have used the internet every day without any restrictions. I have been to every corner of the internet and have watched a serious amount of pornography. I've watched some very weird things, things I couldn't describe and things that no one should ever watch, especially children. I hate myself for watching things I knew weren't good for me. I feel disgusting for the things I have watched. I know my childhood is ruined and that there is nothing I can do to change the past... I'm very sad and I don't know what to do?" Call to Childline from a boy, aged 13[2]

The definition of 'highly effective age assurance' (HEAA) is critical to this measure working effectively and as intended – we discuss concerns we have with Ofcom's current definition in answer to Q.31.

We recommend the guidance gives greater clarity on what services should do to retrospectively identify children on their sites. Research indicates that around a third of children have at least one online account with a user age of 18 or over.[3] Whilst a service may therefore be able to currently claim that it only has adult users, in reality they may have a significant number of children on their site who would go unprotected if they are not identified.

Ofcom must explicitly require services which claim that children are not normally able to access them use HEAA for all users (aside from the exceptions which Ofcom sets out in the guidance). This must include users with existing accounts, as well as new accounts.

[1] Ofcom (2024) Children and Parents: Media Use and Attitudes Report.
[2] 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 used in this response.
[3] Ofcom (2022) Children's Online User Ages Quantitative Research Study.

2. Our proposed approach to the child user condition, including our proposed interpretation of "significant number of users who are children" and the factors that service providers consider in assessing whether the child user condition is met?

Significant number of users who are children

We support Ofcom's approach and interpretation of the legislation in this section.

Ofcom's statement that 'even a relatively small absolute number or proportion of children could be significant in terms of the risk of harm to children' is particularly welcome. Ofcom's approach will have an important impact on reducing harm, as it will ensure that platforms which have a predominantly adult user base, or even a substantive minority of child users, cannot argue that they are below the definition of 'significant'. This is especially important for sites such as Telegram, which children use and poses a significant safety risk, but may claim that they are outside of the Child Safety Duties.[4] Ofcom's approach will ensure comprehensive standards of protection and help reduce harmful content being displaced onto sites which are out of scope of the child safety duties.

It is worth noting that Government indicated this was their intention in the passage of the Act. During the Bill's first Committee Stage, then DCMS Minister Chris Philip noted that the reason for including 'significant number' was to ensure that platforms that either have no children accessing them or pose no risk – e.g. a website on corporate tax – did not face disproportionate regulatory obligations.[5] It is therefore in-keeping with the Act and the Government's intentions to ensure that sites which do have children accessing them and/or do pose a risk are in scope.

Services likely to be accessed by children

Understanding what types of content appeal to children is an important stage for platforms in the Children's Access Assessment. Ofcom rightly recognises that 'children want to engage with services not specifically targeted at them'. We agree with this, and think this could be better represented in the indicative examples Ofcom provides of content that is likely to appeal.

In particular, the guidance should be clearer that content which 'appeals' to children includes both content that is interesting to them and they enjoy consuming, as well as content which they view out of curiosity or due to peer pressure which may be harmful to them. For example, research has shown that children are drawn to conflict, violence, and extreme challenges online – either out of personal interest or because they feel pressure from their peers to engage with this content.[6] It is also well documented that some children, particularly those struggling with mental health problems, will intentionally seek out eating disorder, self-harm, and suicide content.[7]

Examples of 'content that appeals to children' must include reference to all forms of content (including images, videos, texts and other formats) which evidence shows children may seek out which is risky / harmful nature.[8] This includes violence, extreme challenges, and eating disorder, self-harm, and suicide content.

---

[4] Oxford Mail (2020) School raises concerns about child safety on Telegram app.
[5] Online Safety Bill Public Bill Committee. 9 June 2022. Column 313. Hansard.
[6] Revealing Reality (2023) Children's Media Lives. Ofcom; Revealing Reality (2023) Anti-social Media: The violent, sexual and illegal content children are viewing on one of their most popular apps.
[7] NSPCC Learning (2022) Children's experiences of legal but harmful content online.
[8] Draft Child Access Assessment Guidance, p20, table 8.

Children are an incredibly diverse population. There will be a wide range of factors which determine what children seek online that can lead to harm, including on unexpected topics, as demonstrated by the snapshot below.

Services must take an evidence-based approach which utilises internal and external data, to ensure that assumptions about children's experiences do not take precedence over the reality of their online lives.

We recognise the value of the case studies included in the draft Children's Access Assessment (CAA) guidance to help illustrate where the child user condition is likely to be met / not met. It is important to avoid an approach which essentially age-gates large parts of the internet, and so we think it is appropriate to set out where services that do not use HEAA would be out of scope of the Children's Safety Duties.

However, there may be unintended consequences to this approach. A common grooming tactic is for offenders to redirect conversations with children to other spaces, known as cross-platform risk. This can include from public spaces to private channels, or from gaming sites to ancillary chat platforms. There will be scenarios where children are redirected onto services which are not aimed at children – potentially to evade detection, avoid platforms with a greater regulatory burden, and/or further isolate the children. This is also a risk we raised in our response to the Illegal Harms Consultation (p14).[9]

Ofcom should use their information gathering powers to better understand where and how children are being redirected online. This insight should inform further iterations of the CAA Guidance.

Proportionality and impact assessments

The approach taken to the CAA is highly proportionate. Services who know that children are definitely or likely using their services will not be impacted by this stage as they can immediately move to the Children's Risk Assessment. For services that want to argue they are out of scope, there will be additional work required. However, providing they are out of scope, they will then not need to implement the Children's Safety Duties so overall the regulatory burden of compliance will be limited.

Service providers might consider introducing HEAA to block children from accessing their service, if they decide this is easier or more efficient than needing to comply with the Children's Risk Assessment and Safety Duties. This could have significant, negative implications for children's rights to access the online world and make use of digital services – including to learn about the world, to form communities, and for creative expression.

To ensure children's rights are not unduly impacted, Ofcom should use this Guidance to remind services of children's right to safe participation in the online world and the importance of delivering this wherever possible. Ofcom should also commit to report publicly on the impact of regulating the online world, which includes any unintended or unforeseen consequences.

3. Our proposed approach to the process for children's access assessments?

The process for the CAA is appropriate and logical.

[9] NSPCC (2024) NSPCC response to Ofcom's Consultation on Illegal Harms.

We support that Ofcom has stated it will use its enforcement powers if services do not complete appropriate assessments. As part of this, it would be valuable if further information was included about how Ofcom will identify and prioritise scrutinising the assessments of borderline services who have determined that they are not likely to be accessed by children, new services which grow rapidly, and those operating in flagrant breach of the regulation.

We also support that when assessing if a service is likely to attract children, companies are prompted to use evidence from external and independent sources. This is vital for ensuring that CAAs are based on the reality of children's experiences online and cannot be skewed by only using company data.

Volume 3: The causes and impacts of online harm to children

Draft Children's Register of Risk (Section 7)

Proposed approach:

4. Do you have any views on Ofcom's assessment of the causes and impacts of online harms? Please provide evidence to support your answer.

Ofcom's assessment of the causes and impacts of online harms is comprehensive. We would reinforce the importance of understanding the most harmful functionalities, as these must be a priority for services to understand and address through the risk assessment process. From contacts to Childline, consultation with children, and our own evidence review, we know these include algorithmic content recommendations; messaging services; and unwanted connections.[10]

"I'm calling about my 17-year-old son, who was recently sent an inappropriate message on Discord, a social network for gamers. This person, who wasn't known to my son, disclosed how they liked to cut themselves – they then sent pictures of what appeared to be self-harm injuries. I haven't seen these images myself; my son has been reluctant to describe what he saw, beyond saying they were very graphic and he can't get them out of his head. He's also been having trouble sleeping. I'm wondering how best to handle this situation. Is this something we should be reporting to Discord?" Call to NSPCC Helpline from a mother

On specific aspects of Ofcom's analysis, we recommend that Ofcom give greater consideration to virtual and augmented reality technologies. NSPCC's *Child Safeguarding and Immersive Technologies* research found that children were accessing VR pornography, which requires specific consideration, including the way it also present CSA risks.[11]

On abusive and hateful content, we support consideration of how this content impacts children's self-expression. This is something that is particularly important in terms of the impact of misogynistic content on girls.

"You've probably heard about Andrew Tate, the influencer famous for posting sexist and misogynistic content. All the boys in my class talk about his videos, they're so influenced by

[10] Bryce, J. et al (2024) Evidence review on online risks to children. London: NSPCC.
[11] Allen, C. and McIntosh, V. (2023) Child safeguarding and immersive technologies: an outline of the risks. London: NSPCC.

him. They started picking on me and some of my friends because we are girls wanting to become things that 'aren't for women.' It's made me feel like I'll never get into my chosen field considering people like them will be in the future generation. I hate it so much but I know I can't do anything to stop it." Call to Childline from a girl, aged 13

5. Do you have any views about our interpretation of the links between risk factors and different kinds of content harmful to children? Please provide evidence to support your answer.

The NSPCC concurs with Ofcom's interpretation of the links between different risk factors and different kinds of content harmful to children. From our evidence review, we know that the services' choice architecture leads to usage habits that increase children's exposure to online risks.[12]

"I get really nervous when my mum checks my phone. There is this game where you make groups and chat to strangers to join their group so you can level up and I talk to these people a lot. I don't give them personal information, but I have given them my Instagram account details. I am now scared that they are going to send me a message and my mum will see it when she checks my phone." Call to Childline from a boy, aged 14

Greater consideration should be given to the links between illegal and harmful content. It is likely that services will be identifying content which borders illegal and harmful – for example, some suicide content. In these circumstances, Ofcom should set out how services should respond.

We are concerned that services may opt to automatically remove legal sexual content, if it is seen to break their Terms of Service, without checking if it is child sexual abuse material (CSAM). It is vital that CSAM is always identified, removed and reported to the relevant external authorities to ensure appropriate safeguarding steps are then taken, and this must be reinforced in this Code.

Services must also assess the interplay between harms. For example, the FBI have reported that adults are grooming children online to sexually extort them and to coerce them into dangerous acts including self-harm.[13] This snapshot highlights a depressive content forum being used for grooming:

"I've been thinking about stuff that happened to me a few years ago. There was so much going on in my life, I'd just started self-harming and the only place I could escape was on Discord. Some of the people on there were total creeps but it didn't matter who they were, I just needed someone to talk to. There was this guy who was 30 or something. He added me and after chatting for a while, he would ask me to, like, self-harm for him and send pics of it, that type of thing. Mum eventually found out and said I was groomed. At the time, I couldn't really process what had happened." Call to Childline from a girl, aged 14

[12] Bryce, J. et al (2024) Evidence review on online risks to children. London: NSPCC.
[13] Federal Bureau of Investigation (2023) Violent Online Groups Extort Minors to Self-Harm and Produce Child Sexual Abuse Material.

Robust and comprehensive risk assessments must require services to assess the links between illegal and harmful content and to ensure they have clear systems in place which ensure this material is swiftly identified, properly reported, and that children have holistic protections and support.

6. Do you have any views on the age groups we recommended for assessing risk by age? Please provide evidence to support your answer.

The NSPCC supports Ofcom's proposed age categories for understanding children's risk of harm online: 0-5, 6-9, 10-12, 13-15, and 16-17. It is a valuable assessment of how children's experiences of the online world develop as they grow older. It is disappointing that this has not been reflected in Code measures, which we discuss further below.

Ofcom lacks evidence regarding the online activities of the under-3s. Academics Lelia Green, Leslie Haddon, Sonia Livingston, Brian O'Neill, Kyle Stevenson, and Donnell Holloway recently released *Digital Media Use in Early Childhood*, which contains extensive evidence on this group. Important findings to note are that parents found the current 'no screens under 2' guidance unworkable and wanted more guidance on how best to navigate the issue of screen time and online access in early childhood.[14] Parents suggested that guidance could be provided to them in 'just in time' places – for example, a feature in app stores clearly indicating which applications would be suitable for their child to use.

7. Do you have any views on our interpretation of non-designated content or our approach to identifying non-designated content? Please provide evidence to support your answer.

The approach set out for identifying non-designated content is based on a very limited interpretation of the Act. Arguably, the Act can be read as defining NDC as all forms of content which present a material risk to children that services identify through their risk assessments and are not covered by PPC/PC. It does not state that Ofcom alone need to determine what classes as NDC, but should instead be seen as a key way to allow services to tackle all risks on their service, including those which are potentially niche to their platform so would not otherwise be identified Ofcom in a Risk Register / Code of Practice.

We understand that that Ofcom does intend for services to consider other types of harmful content within the definition of NDC, and the categories provides are just two examples of what could class as NDC. However, this is not at all clear from the current description in these Volumes.

Ofcom must be clear that services are required to tackle all non-designated content, including content that services identify through their own risk assessments.

Where Ofcom is identifying new forms of NDC, we are concerned with the level of evidence required. Step 3 of the process looks at the material risk of the harm occurring from potential NDC, with the aim of establishing a relationship between significant harm and specific kind of content. ***We are concerned that this will be highly challenging and urge Ofcom to reconsider this Step in particular.*** Ofcom notes that this will be challenging, but then sets out a limited range of evidence that can used to provide insight into if there is a relationship. Ofcom must ensure they can draw on a wide range of insight from children, services and experts. This

[14] Green, L et al. (2024) Digital Media Use in Early Childhood: Birth to Six. London: Bloomsbury Academic.

must not need to be peer-reviewed research, which can be challenging to undertake with children due to ethical limitations and can be a slow process, and must allow for more informal sources. More broadly, we recommend that this step is based on whether it is likely that harm will occur, and the risk to children if this content is not incorporated as NDC, rather than looking for a direct relationship. This could also be informed by existing evidence from similar harms. The limitation of this approach is already apparent in Ofcom's analysis of body image and depressive content; Ofcom has demonstrated that there is significant evidence of the risk posed by this material, but is still seeking further evidence to support their inclusion.

Evidence gathering for future work:

8. Do you have any evidence relating to kinds of content that increase the risk of harm from Primary Priority, Priority or Non-designated Content, when viewed in combination (to be considered as part of cumulative harm)?

The NSPCC's 2023 evidence review found cumulative, passive exposure to harmful content over time leads to more significant harm.[15] There is a particular risk for children already struggling with their mental health, whether they seek out this content or have it served to them by algorithmic content recommender systems. Illegal or very harmful content is often viewed alongside less serious but still harmful content which contributes to the cumulative impact. Cumulative active engagement with hazardous content, such as active membership of pro-anorexia or extremist communities online, leads to significant and severe harm.[16]

The cumulative impact of exposure to harmful content is clear in the experiences of children, as reported to Childline.

"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

"I've been restricting a lot for the past few months, trying to stay under 800 calories. I've also tried to make myself throw up but it never worked – I just end up choking. This all started during lockdown when I randomly started watching these eating disorder videos. It became a bit of an obsession to watch them. I felt fine at first, but then I looked in the mirror one day and something just snapped and I started hating how fat my thighs and stomach are." Call to Childline from a girl, aged 14

9. Have you identified risks to children from GenAI content or applications on U2U or Search services?

The NSPCC has identified a number of risks that Generative AI (Gen-AI) can pose to children. Often, Gen-AI is exacerbating previously known risks to children; calls to Childline refer to social media alongside concerns about AI. Strong safeguarding measures will be necessary to mitigate these.

[15] Bryce, J. et al (2024) Evidence review on online risks to children. London: NSPCC.
[16] Bryce, J. et al (2024) Evidence review on online risks to children. London: NSPCC.

AI CSAM

Gen-AI is being used by offenders and other children to generate fake hyper-realistic images of child sexual abuse material (CSAM). These AI-generated images can be indistinguishable from non-AI content, making it increasingly difficult for police to identify real children, hindering urgent child protection efforts.[17] Children have contacted Childline, explaining that they are nervous to report AI generated images of themselves, or speak to trusted adults, as they may not be believed when they explain that the images are fake.

"A stranger online has made fake nudes of me. It looks so real, it's my face and my room in the background. They must have taken then pictures from my Instagram and edited them. I'm so scared they will send them to my parents, the pictures are really convincing, and I don't think they'd believe me that they're fake." Call to Childline from a girl, aged 15.

From Childline contacts, we know that this technology is already being used to create images to extort children.

"I was talking to this girl on Snapchat who I 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

Images of children being abused can be used to create new CSAM showing these children, re-victimising survivors of abuse.[18] Gen-AI can also be used to modify CSAM to allow it to escape current detection methods. Finally, Gen-AI is rapidly increasing the speed at which CSAM can be generated[19]; this massive proliferation of AI CSAM can normalise the sexual abuse of children, with the risk that offenders will move from AI CSAM to the abuse of children offline. [20]

The evidence shows that this content is being shared publicly and on the dark web, and we also expect that this is happening privately on end-to-end encrypted communications channels.

Grooming

Gen-AI may provide offenders the tools to enhances their ability to groom children online, and the NCA have warned that they expect some offenders will use AI to groom children at scale through automated engagement.[21] Gen-AI can be used to create fake yet realistic seeming social media profiles and convincing real-time voice and face impersonations. The latter technology is being used for fraud and romance scams.[22] There is a risk that abusers will utilise this technology to approach and groom children at scale.

Bullying and Harassment

---

[17] IWF (2023) Prime Minister must act on threat of AI as IWF 'sounds alarm' on first confirmed AI-generated images of child sexual abuse.
[18] Harwell, D. (2023) AI-generated child sex images spawn new nightmare for the web. Washington Post; McQue, K. (2024) Child predators are using AI to create sexual images of their favorite 'stars': 'My body will never be mine again'. The Guardian.
[19] IWF (2023) How AI is being abused to create child sexual abuse imagery.
[20] Crawford, A. and Smith, T. (2023) Illegal trade in AI child sex abuse images exposed.
[21] Virtual Global Taskforce (2024) Technological Tipping Point Reached in Fight Against Child Sexual Abuse. NCA.
[22] Burgess, M. (2024) The Real-Time Deepfake Romance Scams Have Arrived. Wired.

Deepfake technology can be used as a tool of cyberbullying, with children or adults creating manipulated content that damages other children's reputation, self-esteem, and mental wellbeing. From Childline contacts, we know that this is already occurring.

Misinformation

Gen-AI models will sometimes produce plausible but incorrect answers which they state with confidence – often referred to as 'AI hallucinations.' This misinformation can be harmful to children – for example, the Childline snapshot below shows the result of a child asking an AI bot about mental health.

"Can I ask questions about ChatGPT? How accurate is it? I was having a conversation with it and asking questions, and it told me I might have anxiety or depression. It's made me start thinking that I might?" Call to Childline from a girl, aged 12

Next steps for Ofcom

Overall, many of the risks posed by Gen-AI are the same online risks that we have been working to counter already; Gen-AI is exacerbating these harms. Without appropriate safeguards, children are at risk of being exposed to harmful AI-generated content, harmful contact via AI-assisted grooming, and misinformation from AI hallucinations, where AI is integrated into services.

Ofcom should prescribe that services which provide access to Gen-AI applications, integrate Gen-AI products into their service, or allow Gen-AI content to be shared, comprehensively consider this risks that children could be exposed to via their use of this technology. Given that the above harms are covered by the Online Safety Act, it is the regulated service's duty to mitigate the risks. In line with Ofcom's proposed measure SD1, users should be able to report harmful Gen-AI content to all services. Additionally, services must track and monitor how much content is Gen-AI content compared to individually created, the results of SD1 reports, and take action to proactively understand how Gen-AI is impacting user safety on their service. This will enable it to identify emerging trends and tackle new risks as this technology develops.

Additionally, Ofcom should publish a report which comprehensively identifies the potential risks that Gen-AI poses, explains how platforms are dealing with these risks, and what gaps there are in Ofcom's regulatory powers when it comes to tackling these risks.

10. Do you have any specific evidence relevant to our assessment of body image content and depressive content as kinds of non-designated content? Specifically, we are interested in:

a) (i) specific examples of body image or depressive content linked to significant harms to children,

b. (ii) evidence distinguishing body image or depressive content from existing categories of priority or primary priority content.

The NSPCC strongly supports Ofcom's decision to include body image and depressive content as kinds of non-designated content. We have limited further evidence to add, but note that calls to Childline reinforce that children can view body image or depressive content alongside riskier material.

"I have been searching ways to starve myself. I found a website with loads of tips and it hooked me straight away, which is kinda scary. I've been reciting some of the quotes I saw on there whenever I feel hungry; **I've been drinking loads of water before every meal and also after every few mouthfuls, to try and fill myself up faster**; and I've tried to convince myself that the hunger is a sign of me losing weight." Call to Childline from a girl, aged 17

"I have some concerns about my cousin who's 16. **She's been sharing all these videos on TikTok about self-harm, suicide and depression related stuff**. It's not her in the videos, she's reposting stuff from other people. I did message her to ask if she's ok but she's not replying. I'm worried that she won't get the help she needs if she's not telling anyone." Call to Childline from a girl, aged 18

Young people the NSPCC consulted also raised their concerns with this content.[23] When discussing key drivers of harm to children online, several children raised that posts which 'vent' about mental health, and posts which romanticise unhealthy behaviours are commonly pushed to them through algorithms which can negatively impact their mental health.

11. Do you propose any other category of content that could meet the definition of NDC under the Act at this stage? Please provide evidence to support your answer.

At this stage the NSPCC does not have suggestions for content that could meet the definition of NDC under the Act. We encourage both Ofcom and the service providers to continuously monitor for new trends in harmful content, especially if Ofcom is going to continue with its approach as strictly defining NDC, it is vital that this category is maximised and emerging risks are quickly captured.

For example, in late 2023 the NSPCC became aware of 'underground subliminals', which are videos with 'hidden messages' that claim to influence the subconscious of the viewer. The potential harm of these is clear in the following contact to Childline:

"I've been listening to UG subliminals to become underweight, so I can get attention. UG subliminals are like regular subliminals but like with explicit stuff. They're these audios with hidden affirmations underneath that only your subconscious can hear. They're supposed to send signals to your brain and make things come to life, life if you want to get a mental illness or something. I just feel like nobody notices me and I was thinking if I get more attention, I would be happier. Is that normal?" Call to Childline from a girl, aged 10

Draft Guidance on Content Harmful to Children (Section 8)

12. Do you agree with our proposed approach, including the level of specificity of examples given and the proposal to include contextual information for services to consider?

There are three key considerations which should be incorporated into this guidance.

[23] We consulted the Voice of Online Youth (NSPCC's online youth advisory board), a group of young people in Liverpool, and a group of young people in Watford – totalling 38 children and young people – to inform and shape our consultation response. The age range of the children consulted ranged from 10-17 and was mixed gender. When we reference the young people we consulted throughout this response, it is in reference to these groups.

Firstly, Ofcom have rightly recognised that children and young people are often early adopters of tech. Despite this, the level of evidence required to recognise risks and add measures to the Codes means that new trends and harms are less likely to be identified, and means the Register of Risks is inherently backwards looking, with very little analysis of what harms are likely to emerge and develop.

Secondly, while Ofcom has comprehensively discussed the risks posed by these various forms of harmful content, the NSPCC would encourage greater consideration of the harms caused by features of services, rather than solely the content they host. For example, affirmative approaches on platforms (e.g. 'like' features) exploit children's developmental needs and can lead to negative mental health impacts[24], and algorithms do not only risk pushing harmful content to children, but can also lead them to develop communities with like-minded users which impact their safety and wellbeing. It is vital that features and functionalities are not only assessed in relation to the content that they push, but the behaviours they impact. This has particularly emphasised by the young people we have consulted, who have consistently emphasised that mechanisms which increase their engagement on a platform (such as streaks for regular use and endless scrolling) negatively impact them, even when they are not related to experiencing any other harm. Instead, they view these functionalities as harmful because they are encouraging an unhealthy, overuse of the platform.

Thirdly, the reality is that services will often be assessing content in bulk when tackling harmful content for children on their sites. Whilst item-by-item decisions may be relevant for specific content moderation decisions, it will be vital that services understand the archetypes of PPC and PC and general signals of harmful content, so that they are able to prevent the spread of this material at scale.

To address this, we recommend the guidance:

- ***Includes analysis of how services should identify and assess new and emerging risks to children and incorporate this in the Register of Risks.***
- ***Analyses the ways functionalities and features can cause harm to children independently, and not just the way they exacerbate the risk posed by harmful content.***
- ***Sets out key signals or archetypes which services can use to identify harmful material at scale.***

Volume 4: How should services assess the risk of online harms?

Governance and Accountability (Section 11)

15. Do you agree with the proposed governance measures to be included in the Children's Safety Codes?

In response to the Illegal Harms Consultation we set out our assessment of the Governance measures which are also proposed in this consultation – please see our response on pages 5-7.[25]

[24] Bryce, J. et al (2024) Evidence review on online risks to children. London: NSPCC.
[25] NSPCC (2024) NSPCC response to Ofcom's Consultation on Illegal Harms.

In particular, we would like to reinforce the importance of extending some of the governance and accountability measures to small services as well as large services. Without these processes, small services will be ill-equipped to systematically identify, manage and report on risk. Small services can host significant risk to children 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.

16. Do you agree with our assumption that the proposed governance measures for Children's Safety Codes could be implemented through the same process as the equivalent draft Illegal Content Codes?

Yes, we support this approach.

Children's Risk Assessment Guidance and Children's Risk Profiles' (Section 12)

17. What do you think about our proposals in relation to the Children's Risk Assessment Guidance? Please provide underlying arguments and evidence of efficacy or risks that support your view.

Whilst we broadly support the proposals, we do have some considerations which are set out below.

Measuring impact

One of the suggested indicators for whether the impact of harm on a service is likely to be medium/high is how many child users there are. We welcome Ofcom's decision to base the definition of high and medium on the size of the UK child population. Using general population metrics would downplay the potential impact on children, particularly for sites which are targeted at children and have few adult users, and so this is a positive and necessary approach. Ofcom's analysis also notes that number of children is just one indicator of the potential impact of harm on a site, which we strongly agree with. Exposure to Primary Priority Content can have devastating consequences which should not be underestimated even if there are only a small number of children viewing it.

Evidence inputs

Looking at the definitions of 'core' and 'enhanced' evidence which services are required to use in the risk assessment, we welcome the addition of 'data from content systems' as a core input. 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.

We remain concerned, however, that external evidence, such as the views of independent experts or consultations with users, are absent from the core inputs. In our response to the Illegal Harms Consultation, we set out in detail in answer to Q.7 why this is problematic, and the same arguments apply to the Children's Risk Assessment.[26] We also note that the Children's Access Assessment process does require services use independent evidence. This is positive,

[26] NSPCC (2024) NSPCC response to Ofcom's Consultation on Illegal Harms.

but it is illogical that services are not required to have a similarly strong evidence base for the risk assessment, which is pivotal in determining what safety measures they put in place.

We recommend that the following enhanced inputs are instead categorised as core inputs for large services and for services with multiple risks:

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

In particular, we continue to emphasise the importance of ensuring there are methods for children and young people and people with lived experience to feed into this process. As the groups most directly affected by the operation of these services, they are well placed to provide insight to the specific risks on a service, and how effectively mitigations are working. Ofcom must ensure that these voices are heard through meaningful methods of engagement. We recommend Ofcom consider the Baringa and NSPCC's report on user representation mechanisms to understand solutions which regulated services could implement, and incorporate these options into future risk assessment guidance.[27]

Given the reliance on internal data in this Step, we also suggest Ofcom require services bolster their internal processes for gathering data on risk. For example, Ofcom could consider requiring large companies to hire independent researchers to find risks on their platforms. This would help ensure that the internal data services are using is robust and comprehensive.

Unaddressed risks

In this Volume, it is noted that the Children's Safety Codes 'will not be comprehensive in addressing all risks identified in a risk assessment' for some providers, and Ofcom suggest that services *may* identify additional measures that go beyond the Codes to address remaining risks.

This is a deeply concerning dynamic. We recognise that there are limitations imposed by the Act, and that services are likely to have unique risks that cannot all be captured by the Codes – we discuss this further in answer to Q.25. However, there is significantly more Ofcom can and must do to ensure that services do not identify risks which are left unmitigated.

We have previously argued that Ofcom should include more outcomes-based measures in the Codes of Practice, and not just prescriptive requirements. This approach is particularly crucial for addressing harms which are identified in the Risk Register but not tackled through the Codes. ***Codes must include a requirement that services implement and record their own measures to address all major harms identified in their risk assessment.***

Ofcom should also set out in clearer terms how they will work with services to ensure that outstanding risks are tackled. In particular, if a service identifies a wide range of risks, which are not all addressed in the Codes, and chooses to take no additional action. This approach is in direct contradiction of the Act's requirement for services to be safe by design and provide a higher standard of protection is provided for children. Ofcom should use this provision in the Act to ensure that companies cannot evade identifying all the risks on their service and use all

[27] NSPCC and Baringa (2024) Putting children's voices at the heart of online safety regulation: a study of user representation mechanisms in regulated sectors. London: NSPCC.

powers at their disposal – including information-gathering, using the supervision regime, and generating reputational pressure – to bring about action.

18. What do you think about our proposals in relation to the Children's Risk Profiles for Content Harmful to Children? Please provide underlying arguments and evidence of efficacy or risks that support your view.

The Children's Risk Profiles are comprehensive, well-informed by the Risk Register. The key issue here must be ensuring services use all reasonably available information to understand the nature of harm on their sites; Ofcom must be proactive in understanding where services have marked down their risk.

Our greater concern with this section is that not all risks have requisite Code measures, which we discuss further in other parts of our response.

Volume 5 – What should services do to mitigate the risk of online harms

Our proposals for the Children's Safety Codes (Section 13)

Proposed measures

22. Do you agree with our proposed package of measures for the first Children's Safety Codes? If not, please explain why.

We agree with the measures which have been included in the Children's Safety Codes. However, there are some significant, concerning gaps which we discuss in answer to the next two questions.

Evidence gathering for future work.

23. Do you currently employ measures or have additional evidence in the areas we have set out for future consideration? If so, please provide evidence of the impact, effectiveness and cost of such measures, including any results from trialling or testing of measures.

The development of the Codes will be an iterative process, and new evidence and technological developments will allow for stronger versions in the future. However, we are concerned that there are some significant gaps in this first version that risk undermining the efficacy of the Code and should be addressed from the outset. In terms of the areas Ofcom have set out for future consideration, this concern particularly applies to automated content moderation.

We have previously raised concerns that Ofcom has adopted a very high evidential bar for proving the efficacy of suggested Code measures. We also recommend that Ofcom considers how, as the regulator, they can take a leading role in driving best practice. Both Ofgem and Ofwat have run multi-million pound innovation challenge funds to spur on the development of new solutions in their respective sectors.[28] The FCA and ICO use regulatory sandboxes to help regulated companies innovate and safely test new products without fear of breaking compliance with regulations.[29] ***Ofcom should consider how it can utilise its budget and future income from fines and work creatively to drive the development of innovative solutions which prioritise children's safety.***

[28] Ofwat. Water innovation competitions; Ofgem. Strategic Innovation Fund (SIF).
[29] FCA (2022) Regulatory Sandbox; ICO. Regulatory Sandbox.

Automated Content Moderation

It is a significant concern that there are no measures requiring services use some form of automated content moderation (ACM), particularly for large or multi-risk services.

Whilst the Codes set out what companies must do in response to harmful content, they are much less clear about how this content should be identified in the first place. There is a significant risk that this will enable services, particularly those who are looking to take a 'hands-off' approach to moderation, to avoid putting proactive systems in place. Human moderation alone will not be able to effectively assess whether content is PPC or PC at the scale and speed required to meaningfully prevent children from encountering harmful content.

Without ACM, services will be overly reliant on systems such as user reporting for flagging harmful content. User reporting will be entirely inefficient as a basis for identifying and protecting children from harmful content, with one study finding that for children who had seen harmful content, only half had ever reported a piece of harmful content.[30] Relying on these systems will leave large swathes of harmful content unidentified, it will mean safety is not designed into the service, and it means that children will continue to be exposed to harmful content (as users will need to view the content to make the report) – all outcomes which are out of step with the aims of the Act.

Proactive ACM is widely understood as best practice for user-to-user services, who typically use AI and Machine Learning to scan and filter for content that breaches their terms of service, enabling the automatic removal of the most egregious content, supporting human moderation, and informing prioritisation.[31]

ACM will also be critical for improving moderation in livestreaming. Ofcom have identified the risks posed to children by livestreaming. This was also raised by the children and young people NSPCC consulted, who highlighted that Twitch was a particularly risky platform. They noted that when streamers do 'upsetting or dangerous things', it is disturbing for any child who sees it, but is particularly dangerous because streamers can have positions of influence and may encourage dangerous behaviour in others.

Some examples of services using ACM for PPC and PC include:

- **Meta** report that their ACM systems enable them to detect the majority of the content they remove before it is reported.[32] Recent improvements in these systems have enabled them to more accurately detect harmful content at a greater scale.[33] On Meta services, these systems automatically remove content from a platform, reduce its distribution, or inform human moderation.[34]
- **Yubo** uses a combination of AI and human moderation to reduce inappropriate content and behaviour, with automated systems detecting words, photos or videos which are likely to break their Community Guidelines which are then flagged to their Trust and Safety team.[35] This includes proactive moderation in livestreams, which has enabled

[30] Children's Commissioner (2022) Digital childhoods: a survey of children and parents.
[31] Shah, R. (2023) What Is Content Moderation and What Are Some of Its Best Practices? Sprinklr.
[32] King, J. and Gotimer, K. (2020) How We Review Content. Meta; Meta. Promoting safety and expression.
[33] Meta Transparency Centre. Community Standards Enforcement Report.
[34] King, J. and Gotimer, K. (2020) How We Review Content. Meta.
[35] Yubo. Safety Hub: Safety Tools.

them to monitor for harms including hate speech, the use of drugs and weapons, and discussion of self-harm.

- **TikTok** has automated moderation technologies which scan signals across content including keywords, images, and audio.[36] They report that these systems enable them to rapidly respond to global changes. For example, following the start of the Israel-Hamas war they made changes to their machine moderation models which led to a 234% increase in violative comments removed in Israel and Palestine.[37]
- **YouTube** are utilising ACM to identify and limit repeated recommendations of videos that would be innocuous to view once but could be harmful to young people if seen repeatedly (for example, idealising certain body weights or physical features).[38]

We are not endorsing any particular approach in our response, and indeed the scale of harmful content which children see online indicates that existing industry content moderation systems are not effective enough. However, it is clear that ACM is currently embedded in the moderation practices of services and that it strengthens their ability to protect children. It is a major gap that Ofcom is not recognising this in the Codes by including recommendations on ACM, particularly considering these Codes must identify best practice and push all services to go further if children are to see meaningful changes in their safety online.

The next Children's Safety Codes of Practice must include specific requirements for using automated content moderation tools.

Ofcom should use their information-gathering powers to identify best practice and understand the potential of these tools, including learning from approaches to illegal content, to develop measures which will ensure the effective use of ACM, and address important considerations such as avoiding bias and balancing with input from human moderators.

ACM tools should allow services to tackle harm across their services – including in livestreaming.

Gen-AI

We have addressed future considerations for Gen-AI in other answers.

Impact of choice architecture

We are highly concerned that Ofcom have identified a significant risk to children in the Risk Register – features and functionalities that increase user engagement – and chosen not to address this in the Codes.

Ofcom's own evidence clearly sets out the harms caused by features that increase user engagement, such as infinite scrolling, autoplay features, notifications and alerts. The NSPCC's latest evidence review identified quantification of social activity and popularity as one of the three main features which evidence shows can increase online risk and harm to children.[39]

Children report the damaging impact of 'addictive' services. [CONFIDENTIAL].

---

[36] TikTok. Our approach to content moderation.
[37] TikTok (2024) Our continued actions to protect the TikTok community during the Israel-Hamas war.
[38] YouTube (2023) Building content recommendations to meet the unique needs of teens and tweens.
[39] Bryce, J. et al (2024) Evidence review on online risks to children. London: NSPCC

Choice architecture also interplays with other risks – such as increasing the risk that children will be exposed to unwanted contact from strangers. Calls to Childline show that tools which nudge children to connect with more users can lead them to connecting with strangers, putting them at risk of child sexual abuse.

"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 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

Identified risks to children must be addressed through the Codes of Practice. In this case, it may not be appropriate or practical to recommend individual measures for each functionality that increases user engagement, as they will have different purposes and significance on different platforms. Instead, requiring services take a holistic approach to address the risk posed by these functions will enable them to tailor their approach whilst ensuring risky functionalities do not continue to be rolled out and used unchecked.

In the Codes of Practice, Ofcom must require that user-to-user services assess the combined risk of choice architecture for both illegal and legal harms, and to develop, implement and record mitigations which significantly reduce the risk of these functions.

Appropriate mitigation measures are likely to include turning certain functionalities off for all children, turning certain functionalities off for younger children, and lessening the impact of certain functionalities (e.g. reduced notifications for children).

Children of different ages

Under the children's safety duties in the Act, services are required to effectively mitigate the impact of harm for children in different age groups – recognising that the impact of harmful material will to some extent vary depending on the child's age.

It is concerning that Ofcom's decision in this first Code has been to only focus on mitigating harms that impact all children, and not to differentiate between age groups. We do not think this meets the clear requirements of the Act. In the next Code, it is critical for children's rights that they are supported to have age-appropriate experiences online, which develop as they grow older and gain increased independence. Whilst we recognise there are some limitations to existing technologies, services already carry some level of information about how old their users are and combined with age estimation tools, this should enable sites to better tailor experiences to children of different ages.

We question in particular what consideration has been given to delivering age-appropriate experiences on platforms that allow young users on their services. Whilst 13-17 years-old, the typical age band for children on social media, may seem narrow, many gaming platforms in particular have a much broader age range of children using them.

Young people we consulted raised that there is particularly a gap in terms of provision for young teenagers. They noted that current platforms often seem suitable for younger children (such as YouTube Kids), but are not popular amongst young teenagers. They wanted to see more bespoke experiences for 'children in the middle' who still need safer experiences but would benefit from more freedom than child-focused apps.

There are several examples of services already taking a tailored approach to different age groups, which Ofcom could consider learning from in future Codes:

- **Roblox** operates different levels of chat filtering; all children experience chat filtering, with particularly stringent filters in place for users under 13.[40] They also run an opt-in age verification system, which enables users to verify if they are older than 13 to access enhanced capabilities such as Chat with Voice.[41]
- **YouTube Kids** is designed for younger audiences, with parents able to select content based on their child's age depending on if they are aged 4 and under, 5-8 years-old, or 9-12.[42]
- **TikTok** offers users different settings depending on whether they are 13-15 years old or 16-17. For the former group, measures include that their accounts are set to private by default, other users can't download their videos, and direct messaging is not available. Those aged 16-17 can decide if they have these settings on or off.[43]

There are a wide range of areas where it will be important to ensure children are having age-appropriate experiences. These include:

- Different levels of exposure to non-designated content.
- Different levels of exposure to recovery content, which could mean no exposure for younger children, and limited exposure for older children.
- Increased agency to turn safety settings and functionalities on/off, including functionalities that increase engagement.
- Increased access to certain features for older children.

In the next Code of Practice, Ofcom must give greater consideration to and include measures which ensure children have age-appropriate experiences online. This area would significantly benefit from further consultation and partnership with children and young people and child safety online experts.

24. Are there other areas in which we should consider potential future measures for the Children's Safety Codes? If so, please explain why and provide supporting evidence.

Enforcing minimum age limits

We are incredibly concerned with Ofcom's approach to the enforcement of minimum age limits by services.

It is clear that accessing services below the minimum age limit puts young children at significant risk of both legal and illegal harm – addressing this issue is fundamental for achieving all the safety duties for children in the Act. Calls to Childline show that young children have experienced grooming, been exposed to dangerous material, and bullying and harassment. Ofcom's own data emphasises the scale of this challenge, with half of children aged 3-12 having used at least one social media app/site despite the minimum age requirement of 13.[44]

[40] Roblox. Safety & Civility at Roblox.
[41] Roblox. Verify Your Email Address or Phone Number.
[42] YouTube. YouTube Kids,
[43] TikTok. Teen privacy and safety settings.
[44] Ofcom (2024) Children and Parents: Media Use and Attitudes Report.

"I sent nudes to a friend on Snapchat and now she's asking for money or she'll share it with everyone in our class. I thought the photos were temporary and I didn't realise she could take screenshots of them. My friends have all sent nudes before, so I thought it'd be ok - I'm just so ashamed of how stupid I was for trusting her! My mum wants to tell the police, but I'm afraid this will just make my friend upset and give her a reason to release the pics." Call to Childline from a boy, age 11

"I'm feeling sort of sad because people at my school are saying mean homophobic things on WhatsApp, like LGBT people don't deserve to have their own month. It hurts cos I'm gay myself but no one knows yet. Hearing stuff like this makes it even harder to come out, it's like I'm just a joke to people." Call to Childline from a girl, age 11

The decision not to include any measures specifically requiring services to enforce minimum age limits in the Codes of Practice, and only referencing this through the implementation of Terms of Service creates a significant loophole. In particular, suggesting that there is limited evidence on the efficacy of age assurance technology appears to leave Ofcom open to challenge from services who claim that they have not been able to enforce this element of their Terms of Service because, as Ofcom have themselves stated, there is not the technology available. It is at odds with Ofcom's approach to detecting child users, where highly effective age assurance is required, setting a high bar for services (which we agree with), to then set such a low bar for detecting underage users.

There is a sufficient range of age estimation and age verification processes currently in operation to require services enforce age limits to a reasonable degree of accuracy. 5Rights have identified ten approaches to age assurance which can be used independently or in combination to estimate a user's age, including biometrics, profiling and inference models, capacity testing, and cross-account authentication.[45]

Some services already use these technologies in order to support differentiation between children's ages. Roblox use phone number verification to enable users who are over 13 to indicate their age and access voice chat.[46] Yubo use Yoti's age estimation technology to verify the age of users, which largely relies on a user submitting a real-time photo of themselves.[47] When they rolled out this system, 87% of users were verified on their first attempt.[48] Their age checking system is supported through other measures, such as by detecting users who have created multiple Yubo accounts on one device and using Google image search to identify fake profiles.[49]

Yoti is a notable example of a technology that services could utilise to enforce minimum age standards. Their technology has a 96.99% true positive rate for 6-11-year-olds correctly estimated as under 13.[50] Notably, this is significantly more accurate than reliance on self-declaration of age, which is the approach taken by many services currently and means children as young as five-years-old are able to access them. Praesidio Safeguarding have found that children, young people and parents, are generally receptive and open to AI-based assurance

[45] 5Rights (2021) But how do they know it is a child? Age Assurance in the Digital World.
[46] Roblox. Verify Your Email Address or Phone Number.
[47] Yubo. FAQ.
[48] Yubo. Goal: 100% Age-Verified Users on Yubo!
[49] Yubo. Safety Hub: Safety Tools.
[50] Yoti (2023) Facial Age Estimation white paper.

methods, particularly as they are seen as more inclusive for vulnerable children.[51] These groups were also broadly supportive of the use of biometric data used in combination with AI technologies.

Another consideration for improving age estimation will be how effectively interoperable solutions, such as digital wallets, are utilised. Some children will have access to digital ways of proving their age – such as by having a Young Scots card in Scotland, owning a passport, or using a child-only bank such as GoHenry – which could be incorporated into age assurance solutions.

Given implementing age assurance to distinguish between children of different ages is not currently widespread, there would be a number of crucial considerations to ensure any approach is safe, accessible and privacy-preserving. Research indicates that children may be excluded from using age assurance measures if they do not have hard identifiers (such as formal ID), the process requires parental involvement, if it is too complicated for children with additional needs, and if children self-exclude because they feel their data will not be used securely.[52] Services must consider all these elements in designing appropriate solutions.

Through a combination of age assurance and age estimation technologies, it is possible for services to understand, to a much greater degree accuracy than is currently the case, the age of children using their services. We strongly urge Ofcom reconsiders their approach to this issue in the Codes by taking the following steps:

- ***As a minimum, Ofcom should explicitly recognise that there are a number of options available to services which they could either deploy or learn from to design their own solutions.***
- ***Next, there must be a specific requirement in the Codes that services use an appropriate combination of age estimation and verification technologies to enforce the minimum age of their platform.***
    - ***As part of this, Ofcom should outline likely barriers to access for children who are over the appropriate age and require companies record how they will address these to minimise exclusion as much as possible.***
- ***In time, we would expect Ofcom to incentivise innovation and use their information-gathering powers to work towards recommending the use of highly effective age assurance for this process. Given the importance of stopping young children accessing platforms which are not intended for them, this must be a high priority.***

<u>Tackling bullying</u>

In the Codes of Practice, Ofcom have noted that several measures will not apply to bullying as they would not be effective methods for tackling this harm. We challenge this assumption below. However, we also note that this means there are limited measures in the Codes which will directly target bullying. The effects of bullying on children are serious and can be devastating, in the worst cases resulting in self-harm and suicide.[53] This form of harm must be addressed with the same severity as the other harms covered in the Codes, and we recommend

[51] Hilton, Z. and King, H. Making age assurance work for everyone: inclusion considerations for age assurance and children. Praesidio Safeguarding.
[52] Hilton, Z. and King, H. Making age assurance work for everyone: inclusion considerations for age assurance and children. Praesidio Safeguarding.
[53] NSPCC. Bullying and cyberbullying.

these measures should be significantly expanded upon in the future. Bullying can also overlap with targeted abuse and hate, so the issues we discuss below are relevant for both harms.

One key area which should be addressed in future Codes is the creation of multiple and/or fake profiles. A recurring feature of online bullying includes users creating multiple or fake profiles. This can be done to continue to bully someone after they have been blocked, or to create a fake profile of the person they are bullying to humiliate them. As a result, children who are being bullied can be left feeling incredibly vulnerable and powerless to counter the harassment.

"A couple of years ago, a group of people from school made a fake Instagram account pretending to be me. They basically outed me for being trans and other people would post nasty transphobic comments, saying I was a "freak" and stuff. This fake account went inactive for a while, but recently it came back and the people behind it have started posting multiple things a day. They've even stolen photos from my parents' Facebook to post on the account. I've tried reporting the account since it started and so have many of my friends – but Instagram hasn't done anything about it and I don't know what to do anymore. I just feel so awful, I've been physically shaking from the stress of it all." Call to Childline from a transgender girl, aged 13

"There's this group of girls in my school who bully me for how I look. They're determined to make my life hell online too: they keep adding me to these Snapchat groups where they say nasty things about me. When I block them, they create new fake accounts, and I can't stop being added to new groups. I feel like nobody likes me and that I'll always be the unpopular kid." Call to Childline from a girl, age 11

Young people have also raised with us that fake or multiple accounts should be a priority for tackling. They raised that these accounts are often used to spread hate and to bully other children, and suggested that anonymity enables anti-social behaviour. They also raised that when fake accounts are made of celebrities or role models, they can then promote content or behaviours which are negative and dangerous. At the same time, they noted that second accounts can be an important form of self-expression for children, and so they wanted to see a balanced approach to this issue. They have argued that, as a minimum, it should be made clear when a user has multiple accounts, and if a fake account is reported then it should be banned.

Ofcom have recognised in their analysis of the user support tools that the efficacy of blocking tools is limited where the blocked user can create new accounts to continue to target the victim. The grooming risk posed by fake / multiple profiles pose has also been recognised, and this is something we discussed further in our Illegal Harms Codes (IHC) response (p15).[54] It therefore continues to be a significant gap that Ofcom have not addressed stopping the creation of multiple / fake profiles online, particularly in cases where users have been reported, and we urge that this is included in future consultations.

Alongside proactive content moderation, services should utilise other tools and functionalities to prevent and minimise the impact of bullying and harassment. For example, Meta uses the following tools on Instagram[55]:

[54] NSPCC (2024) NSPCC response to Ofcom's Consultation on Illegal Harms.
[55] We cannot comment on the efficacy of these tools as we do not have access to the necessary information or data; instead, these examples are illustrative of potential solutions. We recommend that Ofcom uses their information gathering powers to evaluate the use and impact of these tools.

- Direct Messaging requests are automatically filtered so that users do not see requests with offensive words, phrases and emojis.[56]
- Users can restrict comments from a certain user, so that their comments are only visible to that person.[57] Meta have reported that over 35 million Instagram accounts have used this feature.[58]
- Comments which are similar to others which have been reported are automatically hidden.[59]
- Users receive comment warnings when they repeatedly attempt to post potentially offensive comments.[60] Meta have reported that reminding people of the consequences of bullying and providing real-time feedback has helped shift the behaviour of some users.

In the next Code of Practice, Ofcom should expand on the measures included to tackle bullying. This must include addressing the way fake and multiple profiles are used to harm children online.

Private messaging

Children experience significant harm on private messaging. Calls to Childline show that children are exposed to Primary Priority and Priority Content harm in private groups – including bullying and abuse, exposure to self-harm and suicide content, and exposure to sexual content.

"A guy I used to date has been spreading false rumours about me on WhatsApp, saying I'm on drugs and I've had sex with loads of boys. There are loads of comments from people I don't even know calling me a 'slag' and a 'crack head'. I try to just ignore it but then I get so paranoid walking round school, wondering what people are thinking about me. It's so stressful and I don't know what to do." Call to Childline from a girl, aged 16

"I got added to a Whatsapp group where people post selfies of other people. Everyone else in it rates how ugly they are and tells them to kill themselves. I'm worried that I'll be identified just from being in the group" Call to Childline from young person, aged 14

"My so-called friend added me to this WhatsApp group chat, where people were saying really horrible things about me, like my parents don't love me and that I should kill myself. At first, I thought it was just some sick joke, but then I realised it wasn't. I kept thinking, why are they saying these things, it doesn't make sense?! I've now blocked this friend, but I don't know how I'm meant to get passed this, like mentally. I feel so hurt, angry and empty right now." Call to Childline from a girl, aged 16

As Ofcom have not recommended the use of any proactive technologies in the Codes, all measures apply to private messaging. In reality, however, private messaging services will have very limited duties under these Codes, despite the harm which occurs on them. They will not be proactively detecting PPC and PC, likely relying on user reporting alone before they take action against content, which means children will continue to encounter significant harm on these

[56] Instagram (2021) Introducing new tools to protect our community from abuse.
[57] Unicef. Cyberbullying: What is it and how to stop it.
[58] Instagram (2021) Kicking Off National Bullying Prevention Month With New Anti-Bullying Features.
[59] Instagram (2021) Kicking Off National Bullying Prevention Month With New Anti-Bullying Features.
[60] Instagram (2021) Kicking Off National Bullying Prevention Month With New Anti-Bullying Features.

services. The user support measures will be of some benefit but do not place sufficient responsibility on platforms to prevent harm.

In particular, there is a risk that harmful content and risky behaviour will increasingly migrate from public spaces to private messaging sites. ***Ofcom must pre-empt this shift and include strong measures for private messaging sites now, otherwise one part of the online world will be safer for children whilst another will put them at continued, or even greater, risk.***

This should include measures which will ensure messaging services take a proactive approach to tackling harm to children. As a minimum, this could include the way services should use the data available to them to identify chats that may pose a significant risk to children – such as by assessing chat names, profile pictures and descriptions, or identifying groups that have been reported by users for causing harm – and blocking children's access to these or closing the groups. This area would also benefit from further evidence gathering by Ofcom.

It is worth noting that the children we consulted called for greater filtering in messages, so that grooming, harmful content, scams, or key words which they chose could be consistently filtered out of their chats. Child have high expectations for what should change in private messaging. Whilst we recognise that Ofcom's powers are limited with respect to private messaging, they must consider how they can go further within the Act.

Developing the Children's Safety Codes: Our framework (Section 14)

25. Do you agree with our approach to developing the proposed measures for the Children's Safety Codes? If not, please explain why.

Whilst we agree with the measures that have been included in the Codes, as with the Illegal Harms Codes we have ongoing concerns with Ofcom's approach to their development. We are pleased to see an increased focus on consulting children and young people in these Codes, and also discuss how this should be built on further below.

<u>Ensuring Safety by Design: Evidence, ambition and outcomes</u>

Ofcom assessment of causes and impacts of harms to children online in Volume 3 is thorough. It is concerning, however, that the full range of causes are then not covered in the Codes of Practice. For example, as we have noted elsewhere in our response, it cannot be the case that particular functionalities are identified as a risk to children, but services will not be required to act on them. To allow services to leave risky features and functionalities unmitigated undermines the basic principle of safety by design which should be embedded by the regulator.

As with the Illegal Harms Codes, a barrier for Ofcom recommending more ambitious and comprehensive measures appears to be the high evidential bar which has been set for proving the efficacy of suggested Code measures.

Not only does this approach mean Ofcom have been unable to address all drivers of harm in the Codes, it also continues to risk removing the incentive for platforms to invest in and rollout ground-breaking safety measures. As platforms will be deemed compliant by only implementing the Code measures (due to the Act making Codes a 'safe harbour'), 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.

It is critical that Ofcom reconsiders both their evidential threshold and their focus on prescriptive measures. The inclusion of outcomes-focused Code measures would provide Ofcom with greater flexibility, enable services to utilise their expertise, and ensure that all identified risks are addressed through the Codes.

Linked to this, we note that Ofcom continues to have a strong focus on the cost of changes when assessing whether to recommend a service. It must not be the case that simply because a service has already rolled out a platform or functionalities which put children at significant risk of harm, they are not required to pay the cost of redesigning a safer service.

In future Codes, Ofcom must include outcomes that services should meet to protect children on their service. This approach should be taken for all risks where specific mitigation measures cannot be recommended. The cost of a measure must not preclude its implementation if it is necessary for safety by design.

Information gathering powers

The Act enables Ofcom to use information gathering powers to inform the development of the Codes of Practice, as well as other guidance. Utilising these powers will be critical for developing robust Codes of Practice. Currently, services take a selective approach to publishing information about the risks on their platforms and their approach to tackling these. Targeted information requests will enable Ofcom to better understand the efficacy of existing mitigations, building up the evidence base, identifying best practice and promoting much needed transparency. We have highlighted areas in our response where this would be particularly useful, including tools which are used to support age-appropriate experiences and prevent bullying.

Under-utilising these powers means Ofcom are creating unnecessary barriers to having a strong evidence base. ***When developing the next Codes, we strongly recommend that Ofcom makes greater use of their information gathering powers to inform their proposals.***

Processes

In both this and the IH Codes, there has been little consideration of the way services can embed processes to ensure their services are safe by design, aside from governance measures.

We recommend that when approaching future Codes, greater consideration is given to how services can introduce processes which ensure children's safety is at the heart of decision-making.

This could include:

- Ensuring Trust and Safety teams work with engineering and design teams so that children safety is consistently considered from an early stage.
- Robust testing of platforms and new features and functionalities, both before they are rolled out and on an ongoing basis, to identify weaknesses in a service's safety measures.
- Ensuring young people's views and experiences directly inform the design of services, such as through youth engagement activities. It is important to note that engagement must be both meaningful and safe. The NSPCC would be pleased to support Ofcom to develop best practice on approaches that platforms could take.

Engagement

As with the Illegal Harms consultation, 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.

We welcome the increased focus on engaging with children and young people in the development of these Codes. Ofcom's significant research programme is important, but it is particularly positive to see the deliberative engagement programme that will be undertaken with children to gather their views on the proposals. The results of this programme must be shared publicly to help organisations understand how Ofcom is being effectively challenged by children and how they are adjusting their approach. It would also be helpful to understand how children's feedback will meaningfully inform decision-making, and if it will be able to be substantively addressed, given there is limited scope for substantial amendments at this point.

Ofcom have rightly recognised the important insight which children and young people have to offer in the development of safety solutions. ***As a next step, this must be embedded in the regulatory regime through the introduction of formal mechanisms to ensure children are consistently able to inform and shape decision-making.***

The NSPCC recently worked with Baringa to understand the features of successful and meaningful user engagement in regulatory regimes.[61] Baringa have demonstrated that, ideally, user representation should be independent, drawing upon expertise, covering the full scope of online safety, well governed and provided in a timely manner. Whilst Ofcom's work for this consultation is a positive step, the criteria of independence and timeliness in particular have arguably not been met – given Ofcom have led this work, and children are providing feedback at a point where limited changes can be made. As noted in the previous section, regulated services would also significantly benefit from engaging with children and young people to better understand the likely success of different safety solutions. The key learnings about user representation from the Baringa report should also inform any future Ofcom recommendations on this topic.

Timings

Ofcom have stated that they intend to publish the final statement and documents in spring 2025. This will be around a year since the start of the consultation, and at least 18 months since Ofcom began to form the Codes.

We are concerned that these time lags are too great. There is a significant risk that the Register of Risks and the Codes will not be sufficiently up-to-date or agile enough to respond to the changing profile of harm. Whilst recognising that Ofcom has certain duties in terms of consultation and Parliamentary approval, we urge that Ofcom considers how to quicken this process. In particular, it is vital that future additions can be done at a much quicker pace to reflect emerging risks to children and young people.

26. Do you agree with our approach and proposed changes to the draft Illegal Content Codes to further protect children and accommodate for potential synergies in how systems and processes manage both content harmful to children and illegal content? Please explain your views.

[61] NSPCC and Baringa (2024) Putting children's voices at the heart of online safety regulation: a study of user representation mechanisms in regulated sectors. London: NSPCC.

Yes, we strongly agree with this approach. It is highly logical to align measures where they overlap between the Codes, and to ensure the strongest protections are in place for children – including to protect them from the most egregious illegal harms online.

27. Do you agree that most measures should apply to services that are either large services or smaller services that present a medium or high level of risk to children?

Yes, we agree that the Codes should take a risk-based approach. The services that pose the greatest risks to children must embed safety regardless of their size. It is important that the governance and accountability measures are in place to support this and ensure all platforms are accurately judging the risk-levels of their service.

28. Do you agree with our definition of 'large' and with how we apply this in our recommendations?

We recognise the benefits of aligning the definition of 'large' between both Illegal and Children's Safety Codes. However, as we raised in our previous response, we think this definition sets a high bar and potentially overlooks the risks some services pose to children.

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 definition overlooks services with a low adult but high child user base.

There are just over 14 million children in the UK. 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.[62] All Volumes suggest that applying measures to services with the highest reach is likely to have the greatest impact for user safety. However, this approach 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.

Limited data availability means it is difficult to know how large many platforms are, but one platform which may not currently be classed as large, but would be large for children, is Kik. Evidence suggests Kik is likely to have less than an average of 7 million monthly users in the UK.[63] However, it is a particularly popular app amongst young people and so would likely be a large service for children. Kik also poses a risk to children's safety, with recognised risks including exposure to harmful content and contact with strangers.[64] 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 platforms in this bracket.

We welcome that in the risk assessment process, Ofcom bases the definition of a high / medium number of users on the UK child population. This child-led approach must be extended to the Codes of Practice too.

We recommend that Ofcom develops a new category of 'large services for children' which is applied to both the Illegal Harms and Children's Safety Codes in the future.

---

[62] NSPCC (2024) NSPCC response to Ofcom's Consultation on Illegal Harms.
[63] Woodward, M. (2024) Kik 2024 User Statistics: How many people use Kik? Search Logistics.
[64] Anderson, D. Parents' Guide to Kik. Common Sense Media.

29. Do you agree with our definition of 'multi-risk' and with how we apply this in our recommendations?

Yes, we agree with the definition of multi-risk services applying to services with a medium or high risk for two or more kinds of content harmful to children.

30. Do you agree with the proposed measures that we recommend for all services, even those that are small and low-risk?

Yes, we agree. It is important that all services have some fundamental safety processes and features embedded into their operation, including on governance and accountability, content moderation and reporting. Services may attract new user bases or develop in a way that leads to increased harm and risk. It is appropriate and necessary that there are systems in place to identify and act when the risk profile of a service changes.

Age assurance measures (Section 15)

31. Do you agree with our proposal to recommend the use of highly effective age assurance to support Measures AA1-6? Please provide any information or evidence to support your views.

<u>Defining HEAA</u>

The definition of highly effective age assurance (HEAA) is fundamental to the efficacy of these measures. We support the technology neutral approach which Ofcom has taken, enabling services to deploy their own or independent systems which are appropriate for their users. However, to ensure consistency and avoid services implementing their own systems which are ineffective in practice, the definition of 'highly effective' is crucial. **By failing to set a specific definition of technically accurate, we are concerned that Ofcom is creating a significant legal loophole in this guidance.**

Ofcom will need to have a benchmark for judging if a service's process is sufficiently accurate. For example, if a service is using a system which has a 30% true positive rate, this would presumably not be accepted as effective and Ofcom would take action against the service. However, because no benchmark is provided in the guidance, services may reasonably be able to push back on Ofcom's decision and argue that it is technically accurate by their own standards. Even if it is a range that is provided, Ofcom must provide some indication of how they will be judging levels of technical accuracy.

Based on the current standards of age assurance, we would suggest that to be defined as highly effective, systems should have a true positive rate for children correctly estimated as under 18 of 95%.

<u>Grooming measures</u>

Services who have a medium-high risk of grooming must be required to use HEAA to ensure that children can access the safety settings and support measures Ofcom proposed in the IHC.

Ofcom recognised in the IHC that the grooming measures would be considerably more effective following the introduction of HEAA. Given the severity of the risk posed by grooming, it is vital that the measures to tackle it are as robust as possible. As the success of the grooming

measures is reliant on identifying child users, it is both necessary and proportionate to require that services with a medium-high risk of grooming are using HEAA.

a) Are there any cases in which HEAA may not be appropriate and proportionate? b) In this case, are there alternative approaches to age assurance which would be better suited?

The Act only explicitly requires that *highly effective* age assurance is used to prevent children from encountering PPC. It is important to avoid a situation where large parts of the online world become age-gated through hard checks. For children, this risks limiting their access to valuable services, making it too burdensome to access sites, and could result in unnecessary data collection.

Services can and should use a range of age estimation measures to better understand who their users are to develop age-appropriate experiences. This may include reducing the prominence of non-designated content on younger users' feeds, and adjusting whether certain safety settings are on permanently or by default. As the definition of NDC in particular expands over the course of the regulation, it may not always be appropriate to require services use HEAA to filter it for children, and instead the estimated age of a child should inform if and how easily they are able to view NDC.

32. Do you agree with the scope of the services captured by AA1-6?

Measures AA3 and AA4

Measures AA3 and AA4 will require services to use HEAA to protect children from PPC / PC if they do not prohibit this content in their Terms of Service. However, there is overwhelming evidence that, currently, platforms do not effectively enforce their Terms of Service in practice. This means that children regularly encounter content which is technically prohibited.

Ofcom's own research assessment of the scale of risk children face online clearly demonstrates that Terms of Service are not consistently enforced. Further key examples of this include:

- At Molly Russell's inquest, Elizabeth Lagone, Head of Health and Wellbeing policy at Meta, recognised that some of the posts and videos which Molly had seen violated Instagram's guidelines, which prohibit the glorification, encouragement and promotion of suicide and self-harm.[65]
- Revealing Reality have shown that vulnerable children regularly view violent, often illegal, activity on Snapchat including fights, beatings, stabbings, sexual assaults, and the sale of weapons. The platform is also used by some children to arrange and amplify fights.[66] This content and behaviour is technically prohibited on Snapchat.[67]
- Research by the Centre for Countering Digital Hate found a community for eating disorder content on TikTok which had reached 13.2 billion views, with TikTok's algorithm pushing this content to child users.[68] TikTok's community guidelines state that they 'do

[65] Milmo, D. (2022) Meta executive apologises over inappropriate content seen by Molly Russell. The Guardian.
[66] Revealing Reality (2023) Anti-social Media: The violent, sexual and illegal content children are viewing on one of their most popular apps.
[67] Snap (2023) Privacy and Safety Hub: Threats, violence & harm.
[68] Centre for Countering Digital Hate (2022) Deadly By Design: TikTok pushes harmful content promoting eating disorders and self-harm into users' feeds.

not allow showing or promoting disordered eating and dangerous weight loss behaviours'.[69]

Requiring the use of HEAA to protect children from viewing PPC/PC cannot therefore be based on whether a service prohibits this content. Many services will already claim that they prohibit this content, or may opt to add this to their Terms of Service to avoid implementing HEAA.

Instead, ***services should be required to use HEAA to prevent children from accessing PPC/PC if they have a medium or high risk of this content on their service.*** This will be a much more effective indicator of whether children are likely to see this material in reality, rather than what is included in the Terms of Service as this is often not reflected in children's experiences.

33. Do you have any information or evidence on different ways that services could use highly effective age assurance to meet the outcome that children are prevented from encountering identified PPC, or protected from encountering identified PC under Measures AA3 and AA4, respectively?

Combining automated content moderation with HEAA will be critical for effectively preventing children from viewing harmful content. These systems will enable services to proactively identify this material and then, depending on whether it is PPC, PC or NDC, and ideally considering the age of the child, take action. For example, PPC would be hidden for all children. NDC may be hidden for younger children, and reduced in prominence for older children.

Young people we consulted raised that there should be greater consequences for users who share PPC and PC, and content from these users should be hidden or downranked for children. They raised this would be particularly beneficial where moderation is more challenging, such as on livestreaming sites, to ensure that risky accounts are not promoted to children to reduce the risk of them encountering harmful content.

Our answer regarding automated content moderation sets out in more detail the practices which services could employ.

35. Do you have any information or evidence on other ways that services could consider different age groups when using age assurance to protect children in age groups judged to be at risk of harm from encountering PC?

There are a number of ways that services could adjust the experiences for older children. This could include hiding PC but with an information about why it has been hidden, to help young people understand more about their online experiences. The young people we consulted suggested that services should make content warnings more prominent on posts and that inappropriate content should be blurred, with users confirming they want to view the material. This could be an appropriate approach for older children viewing PC.

For more detail on our view of age-appropriate experiences, please see answer to Question 23.

Content moderation U2U (Section 16)

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

[69] TikTok. Community Guidelines: Disordered Eating and Body Image.

As we have discussed above, we are highly concerned that there are no proposals relating to using effective automated tools for content moderation. However, we expect services will continue to use a combination of human and automated moderation. Automated systems have a critical role to play, particularly for rapidly identifying illegal material and for effectively moderating vast amounts of content on large services. Systems should be supported by appropriate human input.

Services should be expected to independently quality assure their moderation systems to ensure that they deliver the correct outcome and that they tailor the role that human and automated moderation plays as appropriate.

Measure CM1: Content moderation systems and processes designed to swiftly action content harmful to children

Whilst we do not oppose the measures in this proposal, we are concerned that the focus is solely on how services should respond to content once they become aware of it, rather than introducing proactive, preventative measures which stop PPC and PC from being uploaded and ensure services are able to swiftly identify it.

It is positive that there is some recognition that content needs to be actioned swiftly, however no clear guidance is provided as to what 'a reasonable timeframe' for actioning content is. As services are not required to use automated tools, they may claim a long timescale is required for responding to content, by which point a large number of children could have encountered it.

In particular, there is a strong risk that whilst services may be more proactive in actively checking content which has gained traction and is being promoted via recommender systems, content with a smaller viewership may go unchecked for significant lengths of time. Content promoted via algorithms is also more likely to be seen and receive reports/complaints, again meaning harmful content with less views and so less complaints could go under the radar. Yet vulnerable children are more likely to seek our harmful content, and so they will be at significant risk of encountering material which is particularly dangerous for them to see.

In other sections of the consultation, it is noted that services are not required to have measures that enable them to identify when a user posts or re-posts PPC (Volume 4, p403). It is not clear from reviewing the Content Moderation section why this decision has been reached. There are a number of reasons why it would be valuable for services to identify this. In particular, it would ensure that appropriate sanctions can be taken against users who continuously breach Terms of Services and put other users at risk, and it would ensure that the user support for children could be meaningfully tailored, which is particularly crucial for children at a time of crisis. It would also help services to better understand the nature of risk on their services and ensure they target safety improvements to tackle the greatest harms facing their users. We would expect many large services would have the capability, for example, to identify when a previously banned account or new account with the same IP address reposts content, and to use this to inform moderation.

Children and young people we consulted also highlighted that when users break Terms of Service, and action is taken through content moderation, this information can and should inform safety measures on a platform. For example, they suggested if a live-streamer on Twitch violated the Terms of Service multiple times, services could use this information to inform their recommender systems and ensure this content was not shared with, or automatically hidden from child users. This is a strong example of reasonably available information that platforms

should use to inform their safety tools, but it will only be effective if services are consistently understanding who poses a risk on their sites.

As well as including recommendations for using effective automated moderation tools, we recommend Ofcom strengthens this measure to ensure services are required to take a preventative, holistic approach to content moderation.

This should include:

- ***Providing more detail on what a reasonable timeframe for actioning content is, ensuring it considers how rapidly services should be identifying this content.***
- ***Requiring services to identify when a user posts or re-posts PPC, and have a clear policy for how breaches will inform safety measures.***

Measure CM5: Ensure content moderation functions are well resourced

We agree that services must ensure their content moderation functions are well-resourced, and ensuring this is meaningfully complied with should be a priority for Ofcom. We support the requirements that services consider language expertise and build in flexibility to meet demand to these processes.

As well as this, services should be required 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 to their platforms. For example, if they find that their service is particularly high risk for certain types of PPC, they must ensure their content moderation functions have dedicated expertise to identifying and assessing this material.

Measure CM7: If volunteer moderation is used, provide moderators with materials for their roles.

We support the expansion of content moderation measures to cover volunteers. However, it is highly unlikely that the offer of training materials alone will result in the shift required on sites primarily moderated by volunteers to meaningfully protect children.

One of the most well-known sites moderated by volunteers is Reddit. Between April 2022 and March 2023, Childline delivered just over 40 counselling sessions where the young person mentioned Reddit.[70] The main concern where this was mentioned was mental and emotional health, followed by suicide. The calls showed that children were seeking information on Reddit related to mental health and anti-depressants, and had been distressed after finding subreddits about self-harm and suicide, extreme sexual behaviours, and violence. In the latter situation, children reported frustration that they were not able to report harmful subreddits (only specific posts). Ofcom's research also reinforces the risks posed by Reddit[71], [CONFIDENTIAL].

"I downloaded this app called Reddit and there was an 18+ video which I watched and I was horrified. Some of the images I haven't been able to get out of my head. I know it was dumb to press play but the title was completely different to the video – and some of the comments

[70] Based on analysis of Childline data. Further detail available upon request.
[71] Ipsos UK and TONIC Research (2024) Online Content: Qualitative Research - Experiences of children encountering online content relating to eating disorders, self-harm and suicide. London: Ofcom.

were even worse than the video itself. I'm starting to hate my phone because of it." Call to Childline from a girl, aged 13

"Subreddits are channels I suppose, groups, they're like communities. It's very much tailored to you, so if you wanted to go out and look for it [suicide, self-harm and eating disorder content] you can find it. Unlike YouTube, which are sometimes good at their job of trying to moderate, Reddit isn't as moderated." Ofcom research – insight from a boy, aged 14

Another example is Discord. Young people we consulted raised that because Discord is volunteer moderation, lots of 'creepy' and 'problematic' content is shared on the site, and calls to Childline show that children at risk of seeing harmful content as well as illegal harms on the site.

Platforms which rely on volunteer moderation will need to significantly strengthen their processes to meet the Act's Child Safety Duties of preventing children from encountering PPC. ***We recommend that the next Code of Practice strengthens measures for volunteer moderator. At a minimum, these should apply to large, multi-risk services.*** Measures could include mandatory online training, with moderators required to actively participate in training and engage with materials. Accreditation schemes for volunteers who have completed training could help Ofcom to assess if services are strengthening their volunteer moderation functions.

37. Do you agree with the proposed addition of Measure 4G to the Illegal Content Codes? Please provide any arguments and supporting evidence.

Yes, we support this. The strongest measures and protections for children must all be reflected in the Illegal Harms Codes.

Search moderation (Section 17)

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

We broadly agree with Ofcom's proposed measures, which provide important changes to make search services safer for child users.

Regarding measures SM1A and SM1B, we acknowledge that Ofcom has chosen not to recommend how platforms should determine the severity of content, as platforms should already have risk assessments and internal content moderation policies to draw from, However, we question whether this responsibility should lie with service providers who are likely to take the path of least resistance when assessing harm. Ofcom should monitor the implementation of this measure and assess how services are measuring severity, in order to ensure that these measures are implemented robustly.

40. Regarding Measure SM2, do you agree that it is proportionate to preclude users believed to be a child from turning the safe search settings off?

We support this measure and believe that the manner in which Ofcom proposes implementing it is proportionate and appropriate.

This said, Ofcom should ensure that services are assessed on their approach to the safe search setting itself. If a child is searching for factual support information about (for example) eating disorders, they must still be able to access this. Services should ensure that support services

for topics which come under Ofcom's harmful content proposals do not accidently become blocked by the same measures designed to protect children.

41. Do you consider that it is technically feasible to apply the proposed code measures in respect of GenAI functionalities which are likely to perform or be integrated into search functions?

We consider it to be both technically feasible and highly desirable for companies to apply the proposed measures. Indeed, as these measures have been identified as critical for ensuring search services are safe by design for children, services should only integrate GenAI functionalities into search functions if they are able to comply with these measures. The purpose and power of this regulation is to ensure that safety considerations are embedded into services from the outset, and this is a critical opportunity to ensure GenAI cannot be rapidly rolled out across search services in a way that puts children at risk.

42. What additional search moderation measures might be applicable where GenAI performs or is integrated into search functions?

Where generative AI is integrated into search functions, AI-generated content should be clearly labelled as such and include clear warnings regarding AI hallucinations and the potential for misinformation. This will aid users to make informed decisions about the information that they are presented with. Gen-AI search services should also be expected to use proactive content moderation techniques on the outputs of their searches, to ensure that harmful content is not being shared with children.

User reporting and complaints (Section 18)

43. Do you agree with the proposed user reporting measures to be included in the draft Children's Safety Codes?

a) Please confirm which proposed measure your views relate to and explain your views and provide any arguments and supporting evidence. b) If you responded to our Illegal Harms Consultation and this is relevant to your response here, please signpost to the relevant parts of your prior response.

In our response to the IHC we set out detailed analysis of children's experiences of online reporting systems and the importance of strengthening these tools. Please see our answers to Question 28 on pages 24-26.[72]

Measure UR2: Have easy to access and use, and transparent complaints systems

We support this measure. As we have set out previously, reporting is currently 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.[73] Improving accessibility and transparency is critical for changing this.

The requirement that complainants are able to include context or supporting material with a report or complaint is beneficial. It will support services to understand key risks on their services which should be used to enhance their safety systems, such as the design of

[72] NSPCC (2024) NSPCC response to Ofcom's Consultation on Illegal Harms.
[73] Thorn (2021) Responding to Online Threats: Minors' Perspectives on Disclosing, Reporting, and Blocking.

recommender systems. It is important to note that, depending on the format that users are able to provide this feedback, it may result in services receiving concerning information about children. For example, if users can add context into a free text box and they are reporting concerns about suicide content, this may include sensitive information about the child's own mental health. ***Services should be required to consider how the information they receive from users will be monitored; in particular, it would be beneficial to make users aware if they will not receive a response from the service, and to signpost to other sources of help.***

In this section, Ofcom states that they have not recommended that service providers collaborate with specialist children's organisations when designing complaints processes because this may be over burdensome for these organisations. As far as we are aware, Ofcom have not discussed this decision with children's organisations who could be involved in this. Whilst Ofcom is right to recognise the limited capacity of this sector, it is vital that Ofcom works in partnership and collaboratively with organisations to understand how their capacity and expertise can be best utilised to support the effective implementation of the regulatory regime. During the passage of the Online Safety Act, there was a strong interest in the issue of effective complaints, and it may be the case that some organisations would have opted to be involved in this process due to its importance to them.[74]

As well as complaints mechanisms, there will be a number of other areas where specialist children's organisations could offer vital expertise – such as the way organisations consult with children as part of their risk assessment process, or developing support information for child users. It would significantly strengthen the regulatory regime if the expertise of this sector was formally built into certain processes.

We urge Ofcom to collaborate with specialist children's organisations to ensure their capacity and expertise can be effectively utilised by the regulator and regulated services.

Measure UR2 (e): provide an explanation of whether the service notifies users when their content is complained about, and, if so, what information the notification includes…

We support this addition to the Codes. 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. Young people the NSPCC work with have previously raised that understanding more about a reporting process would encourage them to use it then and again in the future. We previously called for a new measure where users are provided with information about the reporting process upfront, so we welcome this addition.

However, it continues to be an issue that children's identities will not be in some way protected during this process. Research from Thorn, which has found that one of the top reasons children do not report is because they are worried about remaining anonymous.[75] Calls to Childline show that children are worried about the repercussions if others find out about their online experiences as a result of making a report. The below examples are regarding illegal content, but the concerns are applicable to PPC and PC too – particularly if a child is embarrassed, for

[74] UK Safer Internet Centre (2023) Online Safety Bill – How the UK Safer Internet Centre Campaigned for Online Appeals Processes.
[75] Thorn (2021) Responding to Online Threats: Minors' Perspectives on Disclosing, Reporting, and Blocking.

example about viewing pornographic content, or worried about the repercussions if they are reporting someone else, such as someone who is bullying them.

"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.

"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.

We recognise Ofcom's argument that sometimes it may not be possible for providers to guarantee anonymity, for example because it may be clear who has made a report through a process of elimination or if content was only shared with one user. However, ***services can and should still make it clear to the child reporting that they will not share this information with the person who posted the content / the person they are reporting***. Whilst they can note this does not guarantee their anonymity, it would provide some level of reassurance to the child, reducing a key barrier to reporting.

Measure UR3: Acknowledge receipt of complaints with indicative timeframe and information on resolution

Building on our point above, we support the inclusion of measures which increase transparency around reporting and complaints procedures. We continue to be concerned, however, that children and young people will not receive an update on the outcome of their responses, as research indicates that clear explanations of outcomes and next steps are a key way to improve trust in reporting systems.[76] We discuss this further in our IHC response, and urge Ofcom to reconsider this approach.[77]

Offering children the option to opt out of receiving communications relating to a complaint, in case this causes further distress, is rejected by Ofcom on the grounds that they have not seen evidence that this is a problem and evidence shows children want more information about reporting, rather than less. This is an overly narrow view. Children are currently unlikely to raise that they are distressed by information shared in report updates because they rarely receive any follow-up. ***Rather than focusing on adapting ineffective systems, Ofcom should instead reimagine what best practice looks like and use this as the basis for Code recommendations.*** This also illustrates why it is vital to directly consult children and young people in the design of online safety solutions. Ofcom should be engaging children to build a clear vision of what platforms which are safe by design for children would look like.

Trusted Flaggers

[76] Ilk, 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.
[77] NSPCC (2024) NSPCC response to Ofcom's Consultation on Illegal Harms. Q. 28: 5C. Appropriate action – sending indicative timelines.

As with our point above, there is currently less evidence on the efficacy of Trusted Flaggers because services often do not utilise them effectively. It has been noted that Meta, for example, have under-resourced their Trusted Partner Programme which has significantly undermined its potential impact – it is estimated that approximately 1,000 Trusted Partner reports are submitted per month, but it can often take weeks or months for partners to receive a response.[78] However, the very fact that such a large number of reports come from Trusted Flaggers indicates the potential they do have to inform content moderation.

Well-regulated, these systems can provide an avenue for services to enhance content moderation and reporting systems and ensure they are able to prioritise the most dangerous content for children. ***In the next Code, Ofcom should include recommendations for Trusted Flagger systems, including ensuring principles for best practice are included.***

44. Do you agree with our proposals to apply each of Measures UR2 (e) and UR3 (b) to all services likely to be accessed by children for all types of complaints?

Yes, we agree with this proposal. This is particularly crucial for ensuring lower risk services have the reporting mechanisms in place to allow users to raise where there are harms on a platform which can feed back into a service's trust and safety processes. It is important that the burden does not sit with users to identify emerging harms, however, which is why this should be combined with more expansive governance and accountability measures for smaller / low risk services, as discussed in Q.15.

45. Do you agree with the inclusion of the proposed changes to Measures UR2 and UR3 in the Illegal Content Codes (Measures 5B and 5C)? a) Please provide any arguments and supporting evidence.

Yes, we support this. It is vital that the strongest measures are in place for the most serious forms of harm online, including illegal harms faced by children and young people.

As well as this, across both Codes, Ofcom should use their transparency tools to understand the key issues which users are reporting and making complaints about to determine whether these risks are adequately addressed in the Register of Risks and Codes.

Terms of service and publicly available statements (Section 19)

46. Do you agree with the proposed Terms of Service / Publicly Available Statements measures to be included in the Children's Safety Codes?

Measure TS3 (Children's Safety Codes) and New Measure 6AA (Illegal Content Code): Terms and statements for Category 1 and 2A services contain the findings of their most recent children's risk assessment / illegal content risk assessment

We support the inclusion of this transparency measure in the Codes, but disagree that the level of detail provided by Ofcom is sufficient. **Ofcom have provided very limited information as to how services could comply with this measure. This risks services failing to provide detailed summaries of their risks assessments, and instead cherry-picking parts which they are addressing whilst leaving out vital information that requires external scrutiny**.

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

[78] Internews (2023) Safety at Stake: How to Save Meta's Trusted Partner Program.

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]

External actors, including civil society organisations, must be able to assess the steps services are taking to identify and address the harms on their service. This is critical for building in transparency and accountability into the regulatory regime as a whole, and ensuring the risk assessment process is not a closed discussion between service providers and the regulator. The value of this will include civil society being able to undertake targeted research, and raise where their own research and insight contradicts the conclusions of services. It will also allow civil society to better understand the key harms children are experiencing – both in nature and scale – to inform the development of services that respond to these harms, ensuring limited resources are well-targeted.

It is highly likely that, without clear instructions in place, services will continue to bury or underplay evidence of risks, particularly those that they are unwilling to meaningfully address. ***To ensure that this measure has the outcome of improved transparency, we recommend that further details are provided as to how services should comply, including requiring services to:***

- Outline the sources of evidence used to inform their Risk Assessment.
- Detail all the risks and harms they have identified on their service (as part of Step 2 of the Risk Assessment). This should not be limited to the harms and risks Ofcom have identified in the Register of Risks; all risks uncovered should be reported on.
- Detail which Code measure they are implementing, and where they are developing their own solutions (in line with Step 3 of the Risk Assessment).

48. Do you agree with the proposed addition of Measure 6AA to the Illegal Content Codes?

We agree that this measure should be added, but as outlined in Q.46, believe it must be significantly strengthened to ensure it has the intended outcome of improving transparency.

Recommender systems (Section 20)

49. Do you agree with the proposed recommender systems measures to be included in the Children's Safety Codes?

We strongly welcome the inclusion of explicit measures on recommender systems in this Code. As Ofcom's analysis shows, recommender systems pose a significant risk to children online. There is clear and consistent evidence that children's exposure to PPC and PC can often be due to algorithms promoting this material to them – including suicide, self-harm, and eating disorder

content[79]; sexual content; misogynistic content[80]; and violent content[81]. Young people have also raised that they often feel they lack control over what they see online, and that algorithms push content that they do not want to see or engage with.

In particular, it is positive that Ofcom suggests recommender systems should be designed to take a 'precautionary approach' to filtering out potentially harmful content for children. This is appropriate and proportionate, given the significant risk posed by these systems and the importance of tackling cumulative harm. The rights impact of a precautionary approach is also limited by the fact that services are not, at this stage, removing or hiding content, but just ensuring it is not recommended to children where it may pose a risk of harm.

We note that the strength of this measure is in part dependent on the efficacy of a service's content moderation system, and so reinforce the importance of strengthening these measures.

Measure RS1: Recommender systems to filter out content likely to be PPC from recommender feeds of children

Positively, this measure makes it clear that Ofcom expects services to use a wide range of reasonably available information to inform their recommender systems. However, if services do not have effective systems in place to assess content at present, the impact of this measure will be significantly limited. The core aim of the Children's Safety Duties, that children are prevented from encountering PPC, will arguably not be met if services are not able to effectively implement this measure, as recommender systems will continue to push harmful content.

Services must not be able to claim that they do not need to implement this measure because they don't currently have effective detection and moderation tools. The framing of this measure risks entrenching the weak approach of some services, who may not invest in important content identification processes in order to avoid having to use this to feed into their recommender systems – which, as Ofcom have noted, could be costly for service providers.

Services must be required to use effective content identification processes, such as automated content classifiers, if they are operating recommender systems. As a minimum, this should apply to large, multi-risk services.

The implementation process for this measure is currently listed as: (1) use reasonably available information to identify likely PPC; (2) make the signal available to the recommender system; and (3) modify the recommender system to filter out likely PPC for children. A crucial step which is missing from this is ongoing testing of the recommender system, to understand its efficacy in filtering out PPC for children. ***To ensure that services cannot redesign ineffective systems but technically be deemed complaint, a step must be added requiring services test their recommender systems to monitor and report on their efficacy.***

[79] 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; Centre for Countering Digital Hate (2022) Deadly By Design: TikTok pushes harmful content promoting eating disorders and self-harm into users' feeds.
[80] Burgess, S. (2023) Andrew Tate: Controversial influencer pushed on to 'teen's' YouTube Shorts and Instagram video feeds. Sky News; Das, S. (2022) How TikTok bombards young men with misogynistic videos. The Observer.
[81] Family Kids & Youth (2024) Understanding Pathways to Online Violent Content Among Children.

Measure RS3: Provide children with a means of expressing negative sentiment to provide negative feedback directly to their recommender feed

Calls to Childline and work with children shows that sometimes content, which would not be classed as PPC / PC / NDC and could be informative or safe for other children, can be upsetting or worrying. We therefore support this measure to help empower children online, providing them with greater control over what they see whilst not unnecessarily restricting content for all children.

> "I was looking on TikTok the other day and saw that a climate clock has been built and it's displaying six years. Obviously people on TikTok tend to over exaggerate so I'm not sure what's true. My main worry is that no one knows what will happen when the clock hits zero and it's that thought that's really scaring me. My brain also can't seem to comprehend it all and just someone mentioning it makes me start having a panic attack." Call to Childline from a girl, aged 14

Young people we consulted strongly supported this measure. They noted that some sites do this at the moment, but the option to say you are 'not interested' does not consistently come up. They also questioned the impact these systems have, and argued that this new feature must result in changes to their feeds otherwise it will feel like 'speaking into a void' and discourage use.

It is also important that the responsibility for creating safe feeds does not fall to children, however, which is why Measures RS1 and RS2, as well as proactive content moderation, must be as strong as possible.

As with reporting systems, the purpose of 'negative sentiment' functions must be clear and accessible to children and young people. Key considerations will be ensuring children know this option is available and ensuring they know what happens when they use this function (e.g. the content is not removed but the service will hide similar posts; the original poster will not know they have chosen to hide their content). Once this functionality is used across services, it would be valuable for Ofcom to identify best practice. This could be done in conjunction with defining best practice for other child-led safety tools (such as the grooming measures set out in the IHC).

We support Ofcom's point children should be able to '*privately* express negative sentiment on content encountered via recommender feeds'. The private aspect is critical for ensuring that this functionality is not used, for example, to bully other children by consistently publicly disliking their posts.

51. Is there any evidence that suggests recommender systems are a risk factor associated with bullying? If so, please provide this in response to Measures RS2 and RS3 proposed in this chapter.

Online bullying takes a range of forms, and can include both contact and content harms. Whilst contact harm is less likely to be exacerbated by recommender systems (for example, if a user has sent bullying messages to another user), recommender systems could promote material which is being used to bully another user.

Calls to Childline illustrate that content has often been used to bully children online, including edited images and videos, and fake profiles. These do not link specifically to recommender

systems, but it is reasonable to assume that if content is widely engaged with, it makes it more likely that it could be promoted to other users, deepening the risk to the child involved.

"A group of boys at school used deepfake to make a video of me saying I'm gay. They've made fake chat screenshots of me saying I want to do sexual things to them as well. I have questioned my sexuality but haven't come out to anyone, that doesn't stop the bullies though. I want to tell a teacher but it's my word against all these other boys." Call to Childline from a boy, aged 14

"I have been playing a game for a few months and people have started making fun of me. They have made rude photoshopped pictures of me and call me names like "fat" and "ugly." The other people on the game also laugh at me. Some of them have been telling me I should kill myself and be ashamed of my background. I reported it to the moderator who deleted the posts, but the players are still in the game, so they can just do it again." Call to Childline from a child, aged 12

This is a risk which could grow with the increased use of generative-AI tools which can enable children to create content which mocks and humiliates others with greater ease. The risk this poses to children's safety and wellbeing (as demonstrated by the first snapshot above) means it is proportionate to ensure that services prevent this material from being shared by recommender systems, along with other types of PPC and PC.

We strongly recommend that the recommender system measures should apply to platforms with a medium-high risk of bullying.

52. We plan to include in our RS2 and RS3, that services limit the prominence of content that we are proposing to be classified as non-designated content (NDC), namely depressive content and body image content. This is subject to our consultation on the classification of these content categories as NDC. Do you agree with this proposal? Please provide the underlying arguments and evidence of the relevance of this content to Measures RS2 and RS3.

We agree with this proposal. NDC represents content which is unlikely to cause harm in isolation, but when viewed repeatedly or alongside other harmful content poses a risk to children. It therefore is highly logical to ensure that this content is not repeatedly pushed to children by recommender systems, to mitigate the risk of cumulative harm.

User support (Section 21)

53. Do you agree with the proposed user support measures to be included in the Children's Safety Codes?

We agree with the proposed measures, and suggest further ways to strengthen some of these measures below.

Measure US1: Provide children with an option to accept or decline an invite to a group chat

We support the inclusion of this measure. Children experience significant harms on group chats, and it is important that they have greater control over their experiences on these channels. Children should not be required to have to actually click into the group before they are able to choose to decline it – this option should be made available from the earliest stage possible.

This measure must be extended to the Illegal Harms Codes too. Evidence of harms on group chats show that there is often a mixture of illegal sexual content (including CSAM) and harmful content. For example, a BBC investigation found that children in the North East were being added to malicious WhatsApp groups promoting self-harm, sexual violence and racism.[82] Calls to Childline also reinforce that group chats can be used for, and be dedicated to, illegal material.

"Yesterday I got added to a Whatsapp group with hundreds of people in it. Some of my friends got added too and people in it asked our age and if we could send nude pictures. Other people were sending naked pictures and videos, but we all said no that's illegal and blocked them. I've been really anxious since it happened, I feel better for talking to Childline about it" Call to Childline from a girl, aged 11

"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.

To ensure this measure is comprehensive and will be available on services where children at risk from illegal and legal harms on group chats, it must be added to the Illegal Harms Codes.

We disagree with the decision to exclude services with a medium-high risk of self-harm content from this measure. There is significant evidence that children are at risk of encountering self-harm content on group chats, which is also reinforced in calls to Childline. As noted, journalist investigations this year have found group chats which promote self-harm to young children. As part of this investigation, it was noted that schools had warned that a WhatsApp group was encouraging self-harm with a points scoring system.[83] The FBI have warned that criminals are using group chats to target children and extort them into recording acts of self-harm and producing CSAM.[84] Ofcom's own research showed that young people with lived experience of self-harm often create networks or communities via group messaging, which could be a site where harmful content is then shared.[85]

"I keep blocking these group chats that encourage me to self-harm, but they keep coming back. People in the group get angry if you're not harming when they tell you too; it really gets inside my head when I'm trying hard to stop." Call to Childline from a girl, aged 16

To suggest that children do not risk encountering self-harm content on group chats is inaccurate. It also fails to recognise the way harm is likely to migrate from public to private spaces as the regulation is further implemented – something which should be pre-empted

[82] Downs, J. and Lindsay, M. (2023) Nine-year-olds added to malicious WhatsApp groups. BBC News.
[83] Brady, J. (2024) The chilling WhatsApp group spreading malicious content to nine-year-olds. Mail Online.
[84] Federal Bureau of Investigation (2023) Violent Online Groups Extort Minors to Self-Harm and Produce Child Sexual Abuse Material.
[85] Ipsos UK and TONIC Research (2024) Online Content: Qualitative Research - Experiences of children encountering online content relating to eating disorders, self-harm and suicide. London: Ofcom.

through the Codes. ***Ofcom must extend this measure to services with a medium-high risk of self-harm or suicide content.***

It is important to note, as discussed in Q.24, that strengthening measures to prevent harm in private messaging will be critical to addressing the risks posed by group chats. Whilst this measure is welcome, much more comprehensive measures are required to tackle the root of these harms.

Measures US2: Provide children with the option to block and mute other users' accounts, and US3: Provide children with the option to disable comments on their own posts

We support these measures. They could be strengthened by expanding it to similar functionalities which have been noted by the young people we consulted as posing similar risk to children. This includes enabling children to turn off tagging and sharing of their posts. They argued that as well as having these specific options, accounts should be able to easily turn off all of these features which can lead to increased contact from other users, which they raised would be particularly useful if a child is being bullied or harassed on a site.

As Ofcom looks to implement more age-appropriate experiences, it may be appropriate to have some of these options consistently on for younger children, and on by default (but with the option of turning them off) for older children.

Children and young people we consulted also highlighted that there are other tools that this could be extended to. For example, it was highlighted that on gaming, it would be more useful to have the option to turn off voice chat for other users. They also wanted to be able to turn off sharing of their content. We strongly recommend that Ofcom extends these settings to other mechanisms, and in particular considers which mechanisms on non-social media platforms might be useful to include.

Measure US5: Signpost children to support at key points in the user journey

We agree with the points in the user journey when children will be signposted to support. There are two other points that Ofcom should include for signposting to support.

If children post harmful content, for example about self-harm or eating disorders, it may be an indicator that they are vulnerable and at risk. Ofcom's research has found that many online users with lived experience of self-harm and eating disorders had shared content that would be classified as harmful online.[86] In these cases, if content is removed without follow-on support, it risks further isolating the child. This is therefore a crucial point where services can signpost vulnerable children to further support.

Another point was suggested by a young person we consulted. They noted that if comments are found to be abusive and violate a platform's Terms of Service, these comments might be removed or hidden. They raised that the user who has been impacted by the comments does not receive follow-on support, and argued this is an important moment to be signposted to support and resources.

We recommend children are also signposted to support if:

- ***Their content is removed from a site due to it being identified as harmful***.

[86] Ipsos UK and TONIC Research (2024) Online Content: Qualitative Research - Experiences of children encountering online content relating to eating disorders, self-harm and suicide. London: Ofcom.

- ***They have been impacted by content which has broken a platform's terms of service.***

It is recommended that if a platform wants to signpost to a non-public support service or helpline, they should obtain permission from this organisation. We welcome the inclusion of this, and also challenge Ofcom to consider how they can use their resource and capacity to bolster the organisations who will be providing vital support for children who experience harm online. ***In the future, we recommend that Ofcom redistributes part of any income generated from enforcement fines for breaches to the Child Safety Duties to help resource these vital services.***

Measure US6: Provide age-appropriate user support materials for children

Our response to the IHC covers the topic of user support information for children which we also ask Ofcom to consider for this consultation.[87] We make a number of detailed points in that response which must also apply to this user support information, including the importance of:

- Developing messaging which appropriately balances the need to be child-friendly, engaging, empowering, non-victim blaming, and transparent.
- Presenting support information in engaging formats.
- Engaging directly with children or representative groups in order to develop messaging that is useful and accessible.
- Providing guidance to help them in developing effective and age-appropriate information. Ofcom could produce this guidance themselves or commission it.

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.

We also reinforce the importance of consulting children and young people in the development of these material. Children will be well-placed to inform what materials will be impactful and useful, and what is likely to be discounted. Ofcom should consider how the establishment of user voice mechanisms could inform best practice guidance on the development of support materials.

Search features, functionalities and user support (Section 22)

54. Do you agree with our proposals? Please provide underlying arguments and evidence to support your views.

We support proposed measures SD1 and SD2.

We recommend that Ofcom should go further in SD2 to require services coordinate with any crisis prevention hotline or similar service that they intend to direct users towards, to ensure that sufficient capacity is in place to cope with demand and that all children receive the support that they need, when they need it. We recognise Ofcom's point that search services in scope

[87] NSPCC (2024) NSPCC response to Ofcom's Consultation on Illegal Harms. Questions 31-34.

largely already have this measure in place so the likelihood of increased traffic is limited. However, it is a very limited burden on search services and will ensure that they are consistently directing to services which have the capacity in place to support users. This would also ensure the approach is aligned with Measure US5 for user-to-user services.

Ofcom note that they have decided not to recommend that prevention and support information is provided for other types of PPC and PC – but that they will likely consider the role of supportive resources for harms such as intimate image abuse and controlling and coercive behaviour in the future guidance on protecting women and girls online. We are concerned with this approach. Because the guidance does not have the same 'comply or explain' requirement that the Codes of Practice do, including measures in here which Ofcom have recognised could be included in the Codes unnecessarily downgrades these measures. The guidance should be used to support services to understand how harm to women and girls is manifesting on their services, and how they can go further than the Codes to address this. It should not be used to recommend measures which are best suited to Codes of Practice, and we recommend Ofcom reconsiders this approach.

57. Do you consider that it is technically feasible to apply the proposed codes measures in respect of GenAI functionalities which are likely to perform or be integrated into search functions? Please provide arguments and evidence to support your views.

We consider it to be both technically feasible and highly desirable for companies to apply proposed measures SD1 and SD2. Please see our response to Q.54 for our extended thoughts on these measures.

Combined Impact Assessment (Section 23)

58. Do you agree that our package of proposed measures is proportionate, taking into account the impact on children's safety online as well as the implications on different kinds of services?

We agree that these measures are not overburdensome. In places, they are not proportionate – in that they do not sufficiently respond to the scale of harm experienced by children. This is particularly seen in the absence of measures to ensure services remove underage accounts from their platforms, and the failure to address all identified risks in the Codes of Practice.

Whilst the Act recognises the importance of proportionality, its core aims (set out in Section 1) make clear that regulated services are expected to be safe by design, and to afford a higher standard of protection to children. Ofcom must go further in the next Code to deliver on these objectives.

Annexes

Impact Assessments (Annex A14)

60. In relation to our equality impact assessment, do you agree that some of our proposals would have a positive impact on certain groups?

Yes, we agree with this assessment. Ofcom's ongoing consideration of how their proposals will impact children with protect characteristics is critical for ensuring that all children benefit from this regulation.

One area we will continue to advocate for is strong measures to tackle the (illegal and legal) harms experienced by girls online. Girls disproportionately experience harm online. For example, we know through contacts to NSPCC's Childline that the proliferation of misogynistic content online, from individuals like Andrew Tate, is shaping boys' attitudes and behaviours, causing further harm to girls in school and at home. Teachers have also highlighted that the increase in misogyny online is leading to higher rates of sexist behaviour from boys in school.

Whilst the measures in the Code will help target the content that children see, they will also be impacted by adults who can see this content. For example, those who are just over 18, as well as older adults, whose views are shaped by this content. Research by CCDH found that misogynistic and incel communities also pose a child abuse threat, with paedophilia tolerated or even promoted in some incel communities.[88]

"I've just had a massive row with my brother. He has been idolising and speaking about this creep Andrew Tate. My brother knows that I think this guy is absolutely vile but nothing I say or do will change his mind. I've tried talking to my mum about it, and all she did was tell my brother to stop watching his videos, which of course he ignored. I don't think my family realise how damaging Tate is to society, and I can't believe people like my brother look up to him."
Call to Childline from a girl, aged 12

Ofcom's future guidance on harms to women and girls will therefore be critical in introducing holistic protections which tackle risks to girls in the online world – including risks which do not come from content girls see directly, but which are facilitated by the wider online ecosystem. Ofcom must ensure this guidance remains and priority, and we look forward to working with them to ensure it delivers for girls.

The NSPCC is the UK's leading child protection charity with over 130 years in experience safeguarding children from harms. A driving force in the passage of the Online Safety Act, we are committed to ensuring every child is safe online.

We have significant knowledge and expertise, based on a strong research and evidence base and direct work with children, and are committed to using this to advocate for the development of a strong, ambitious, regulatory framework which centres children's experiences and tackles the full range of harms children experience online.

To discuss the NSPCC's response to Ofcom's Protecting Children from Harms Online Consultation further, please contact Rani Govender (Policy and Regulatory Manager) – rani.govender@nspcc.org.uk.

[88] Centre for Countering Digital Hate (2022) The Incelosphere: Exposing pathways into incel communities and the harms they pose to women and children.